NOVA IMS Information Management School

MGI

Mestrado em Gestão de Informação
Master Program in Information Management

Analysing data mining methods in sports analytics: a case study in NHL player salary prediction

Stepan Mincev

Dissertation presented as partial requirement for obtaining the Master's degree Information Management

NOVA Information Management School
Instituto Superior de Estatística e Gestão de Informação
Universidade Nova de Lisboa

NOVA Information Management School
Instituto Superior de Estatística e Gestão de Informação
Universidade Nova de Lisboa

ANALYSING DATA MINING METHODS IN SPORTS ANALYTICS; A CASE STUDY IN NHL SALARY PREDICTION

by

Štěpán Minčev

Dissertation presented as partial requirement for obtaining the Master's degree in Information Management with a specialization in Knowledge Management and Business Intelligence

Advisor: Mauro Castelli

March 2021

Acknowledgments:

This project could not have been achieved without the support of countless members. I would like to thank my friends from Portugal, United Kingdom and from back home, Czech Republic, along with my supervisor Professor Mauro Castelli for providing valid feedback which ultimately led to the completion of this project. More importantly I would like to thank the entirety of my family, for their continuous support. My brother Krystof who helped guide me from the offset and my wonderful parents Irena and Michal for their ongoing support throughout my academic endeavors. This work truly would not have been possible without the overall group effort, and the endless support I received throughout my journey to the end, thank you.

Abstract:

The deployment of Internet of Things has become a systematic phenomenon around the world, leading to the exponential growth of data and data analysis practices. This particular growth is being seen within the sporting industry as new hardware and software are continuously being developed for home and professional use. Though there are several use cases of effective data usage within elite sports, there remains the notion that professional sporting organizations should expand their resources to fully cease the possibility of competitive advantage, through effective data mining techniques. This project conducts a comprehensive analysis of extensive open-sourced NHL data, utilizing SAS's established SEMMA process. Through the SEMMA process, this project yields a predictive data-mining model, designed to predict future player salaries. With player salaries within the NHL steadily increasing, reaching upwards of $10millon per year, a predictive model with an overall average error of $150,000 and Mean absolute error of $870,000 can grant team's unique knowledge, which if used effectively within the NHL, can lead to superior decision making. Though there remain limitations due to unquantifiable variables linked to a player's psychology, as a whole, concrete deductions show that if effectively analyzed, sporting organizations have the power to leverage data to develop a competitive advantage. Our research indicates concludes that organizations pushing towards developing an established data science department are increasing their odds of winning.

Keywords:

Data mining; Big Data; SEMMA; Decision making; Elite Sports

Index

Table of Figures

List of Tables

List of Abbreviations

NHL	National Hockey League
NFL	National Football League
SEMMA	Sample, Explore, Modify, Model, Assess

1.0 Introduction

The 21st century has seen a rapid surge of technological advancements, particularly the development of data gathering methods and their importance within decision making. The importance of data has led to the term big data which in 2001 was defined by Gartner as Data and Analytics of high volume, velocity and varied "information assets which require cost-effective innovations for information processing to ensure enhanced insight, decision making and process automation" (Gartner, 2018). Despite the recent interest in big data, the term was initially conceptualized in the 1960s during the origins of data processing centers, however merely within the 21st century are organizations and entire industries beginning to understand the potential behind various types of data. The notion of big data has gained such traction that many academics identify it as having more value than oil, with The Economist publishing a story titled, "The world's most valuable resource is no longer oil, but data" (Bhageshpur, 2019).

The scientific development of wisdom through the established hierarchy of data processing; conversion of raw data into information, information into knowledge and lastly knowledge into wisdom; has a vast array of use cases ranging throughout industries (Rowley, 2007). Growing industrial sectors include healthcare where data is used for predictive analysis such as with cancer research, to fraud detection whereby data mining techniques are used to evaluate documents and their validity. The spectrum of data mining has seen rampant growth since its earliest use by Piatetsky-Shapiro who coined the term (Piatetsky-Shapiro & Parker, 2018). The term data mining has since been further established, with the concrete definition being "Data mining is a process which finds useful patterns from large amounts of data" (Kalyani M Raval, 2012). Despite these examples and the rapid interest in data, many industries are still struggling to accurately implement data mining alongside big data to effectively enhance their decision-making processes. In particular, the realm of sports retains the characteristics which can further be enhanced through concrete data mining techniques.

1.1 Background

The sporting market has substantially grown into a multibillion-dollar industry. In North America alone the sports market is estimated to be worth over $75 bn, with forecasting predictions portraying an annual growth of 1.3%, surpassing the $80 bn mark in 2022 (Statista, 2018). Furthermore, throughout the years, the introduction of statistics has been used to track every slight detail throughout a sporting season, with several strategies being developed solely from statistical data analytics alone. Within the National Football League (NFL) the use of clustering techniques saw the norm shift from kicking the ball down the

field to a statistically driven "4th-Down Bot" during a 4th down (Lopez, 2020). However, there is a continuous need for further data analysis within several sporting leagues, as certain aspects of each game are still left to old ways.

The key to winning in team sports has always been based on the relative superior effectiveness, coaching, and front offices. With top teams heavily relying on key specialists who traditionally relied on an instinctive "gut" feeling or observance of traditions to deploy decision-making strategies. However, this notion of a "gut" feeling saw a dimensional shift when Oakland Athletics' General manager (Billy Beane) prioritized the prioritization of statistics and data to determine decisions within the game of professional baseball (Steinberg, 2015). This paradigm shift has seen all major professional sports teams expand their analytics department; however, certain leagues are still relying on old scouting and drafting techniques; any team which does not successfully implement modern analytical techniques is at a competitive disadvantage.

The latest phenomenon has been seen with the growing success of Liverpool. Led by Ian Graham, the Liverpool Football club, identified the use of data to decide upon key momentous decisions. One major decision included the signing of a new manager Jurgen Klopp in 2015. In doing so they began to mix the realms of intuition and data to incrementally develop superior tactics and signings (Schoenfeld, 2019). Liverpool's success story greatly indicates the tremendous power, data has in empowering decision-making within sports.

The prevalence of data has impacted how managers, scouts, or coaches assess individual teams or players simply by reading one set of statistics. The presence of big data has made a tactical analysis based on observational data obsolete, particularly due to the development of new variables based on contextual information (Memmert, 2016). Largely due to the advancement of technology, through sensory technology, various sporting teams have access to new equipment leading to previously unattainable sports-related information the likes of physiological performance, technical abilities, or team tactical behavior. A prime example is highlighted within the National Basketball League which has begun to utilize tracking technology to generate new information about each basketball game performance (Jaime Sampaio, 2015).

Big Data, although not having a direct definition, is described through its characteristics, namely Volume, Velocity and Variety (Aisyah Mohd Noor, 2015). Big data has been embraced thin sports; the volume of data produced is exponentially increasing and the speed at which novel data is being generated is rising.

Such instances can be seen within the world of football; the presence of different types of data is seen with XML files, video recordings, notational meta-data, or the likes of health records; within the Bundesliga solely tracking data amounts to over 400 gigabytes of data per season(Pääkkönen P, 2015). Notably, the notion of big data within sports requires specific analytical actions, as no longer can intuitive behavior be utilized in recognizing patterns within statistical data.

1.2 Relevance

Despite the growing uptake of data within sports, certain aspects are still largely performed through personal interaction and intuition. One concrete environment which still requires modern-day evaluation is salary analysis; whether a team is better off paying a high fixtured salary for one player or divide that fixture between several players. This notion is highly prominent within several American-based leagues which contain salary caps and must optimize their active team roster, in a manner by which the team does not solely expend its salary on one player.

The National Hockey League (NHL) consists of 31 teams spread across the USA and Canada. The salary cap for each team is decided upon annually, with a board generating a margin by which each team needs to operate; within the current season, teams were allowed to spend up to $81.5million on their minimum of 46 players. With a vast talent pool and an annual draft, teams must decide which players to keep for the upcoming season and which players they will not be able to adhere towards. Additionally, there is a large problem with player salaries, as they can drastically change from one season to the next, and often teams may be required to draft a team that meets regulatory standards. A prime example is seen with the Arizona Coyotes, who currently are determining which players to keep, such as their key playmaker Taylor Hall (CapFriendly, 2020). To better understand player salaries, and their future predicted values, thus allowing for superior decision making, this dissertation will develop a predictive model utilizing an open-sourced NHL database.

The chosen database contains information on over 1000 NHL players through a period of 10 seasons, spacing from 2007 to 2016. Two datasets were combined: one representing key player statistics such as Goals scored and Assists and the other representing player salaries. Through the process of data mining, this project will aim to devise a model which can accurately predict player salaries, thereby mitigating any future salary complications. All data files have been attached alongside this document.

1.3 Research Objectives

With the emerging topic in sports academic literature, the role of data mining has been largely discussed regarding its role in improving decision making. Hence, this project seeks to investigate key areas of interest.

This project seeks to, firstly, conduct an analysis of current literature within the field of machine learning and data mining. The literature review will analyze current academic papers regarding the concept of data mining techniques within the sports industry today and the future use of concrete processes.

Secondly, a Jupyter notebook containing python code will be developed yielding a predictive model for future player salaries. By designing a model capable of predicting player salaries, teams would be able to construct an overall player base that is worth the money spent; as there might be players who are predicted to have a high salary based on their data, however, yield a low actual salary or vice versa. Hence, the designed model could entail certain players' values are either too high or too low, thus potentially constructing tacit knowledge which specific teams could take advantage of, to help their seasonal standings.

We conclude this project by describing the overall insight regarding the future of data modeling in sport, based on academic literature and a designed model.

Objectives can be summarized as follows:

Literature Review:

- Identify key definitions based on academic literature
- Identify key motivators behind the use of data mining application
- Identify the use of Internet of Things
- Identify data modelling techniques based on academic literature

Methodology:

- Formulate a predictive data model based on 3rd party data

Data Analysis:

- Analyze constructed data model and how it can be used in decision making

Discussion and Conclusion

- Summarize key findings
- Recommendations for further studies

1.4 Project Outline

This project consists of V chapters, below is a summary of each chapter.

Chapter II – Literature Review

The Literature review comprises of a deep insight into background research. In particular, the literature focusses on academic insights into key contributing factors, these include the concept of data mining and big data, the importance of data mining practices, and other insights regarding sports data science analysis.

Chapter III – Methodology

The methodology chapter identifies central research methods used within this project, primarily focusing on research design. This section will identify key practices used in order to achieve a designed model for data analysis.

Chapter IV – Data Analysis

Data Analysis proceeds to analyze 3rd party data gathered from the National Hockey Leagues open-sourced data base. Data used will be used to identify key patterns, in aims of developing a model which can to a certain degree predict player worth for the upcoming season. Furthermore, patterns within the data will form a base for recommendations.

Chapter V – Discussion and Conclusion

The final chapter focusses on summarizing information, through possible limitations, the impact of gathered findings and how these findings can be implemented within a real-life situation to enhance decision making.

2.0 Literature Review

This Literature review aims to provide a necessary foundation on data mining and the prevalence of data surrounding the sporting industry. Furthermore, key areas of academic literature are identified to obtain a detailed viewpoint. This chapter will grant context of existing literature and the gaps within the relationship of works, forging a case for this project.

2.1 Evolution of Data into Big Data

Data has been a breakthrough technological development over the recent year. Driven by the rapid surge of social media and the Internet of things (IoT) phenomenon. Technological advancements have led to the increasing amount of data being available throughout a vast array of industries, focusing largely on conceptualizing and the extraction of knowledge from numerous amounts of data sources (Lee, 2017). Data and its evolution to big data is an emerging topic of interest within the field of Information systems, management, social sciences, and varying degree of other fields. The phenomenon has accredited its widespread adoption namely due to the rise of social media, mobile devices, sensors and the overall exponential growth of IoT's (Patrick Mikalef, 2017). The IoT is a novel paradigm shift within the world of Information technology, the concept grasps the idea of the internet as a network of networks consisting of millions of private, public, academic, business, and government networks, of local to global scope, which is joined through an array of electronic technologies (Somayya Madakam, 2015). The idea of IoT's greatly extends merely the growing online presence of users, which has surpassed four billion as of 2020 (Statista, 2020), however includes the growing array of devices that can obtain and monitor various data sources. To first identify the various use cases of data sources, it is critical to identify an in-depth definition of the notion of data and big data, as various academic authors have stated their own concrete definitions each with a varying degree of similarity. The history behind data and the idea of gathering and storing large amounts of information date to the early 1950s with the introduction of the first commercial mainframe computer. Since the world of IoT has seen several evolutionary steps, enabling the world of data to develop into big data.

2.1.1 Big Data

To set a starting president, it is necessary to conclude an overview of big data and how the term has been defined in past studies, including which attributes are fundamental to the notion. Over the last decade, several definitions have arisen in efforts to differentiate the phenomenon of big data, from other conventional data-driven and or business analytics-driven advances. The more notable definitions of the

term utilize the already mentioned three V's (volume, velocity, and variety) with the aims of properly aligning the term with modern uses. Table 1 depicts the varying definitions set by several academics in chronological order.

Author	Definition
(Russom, 2011)	Big data involves the data storage, management, analysis, and visualization of very large and complex datasets.
(White, 2011)	Big data involves more than simply the ability to handle large volumes of data; instead, it represents a wide range of new analytical technologies and business possibilities. These new systems handle a wide variety of data, from sensor data to Web and social media data, improved analytical capabilities, operational business intelligence that improves business agility by enabling automated real-time actions and intraday decision making, faster hardware and cloud computing including on-demand software-as-a-service. Supporting big data involves combining these technologies to enable new solutions that can bring significant benefits to the business
(Schroeck M, 2012)	Big data is a combination of volume, variety, velocity and veracity that creates an opportunity for organizations to gain a competitive advantage in today's digitized marketplace
(CK, 2014)	Big data consists of expansive collections of data (large volumes) that are updated quickly and frequently (high velocity) and that exhibit a huge range of different formats and content (wide variety)
(Akter S, 2016)	Big data is defined in terms of five 'Vs:' volume, velocity, variety, veracity, and value. 'Volume' refers to the quantities of big data, which are increasing exponentially. 'Velocity' is the speed of data collection, processing and analysis in real-time. 'Variety' refers to the different types of data collected in big data environments. 'Veracity' represents the

	reliability of data sources. Finally, 'value' represents the transactional, strategic, and informational benefits of big data

Table 1: Definitions of big data

In addition to the several academic definitions, numerous scholars have deducted several different concepts when defining big data. A key conceptualization includes the acknowledgment of veracity as included by Akter et al., the degree of trust, as it is imperative to make cognizant decisions and derive business value. Additional dimensions to the current 4V's include two dimensions conceptualized by Seddon and Currie (Seddon JJ, 2017) these being variability and visualization. "Variability concerns how insight from media constantly changes as the same information is interpreted in a different way, or news feeds from other sources help to shape a different outcome"; and "Visualization can be described as interpreting the patterns and trends that are present in the data" (Seddon JJ, 2017). In its entirety, the most recent and recognized academic definition of big data is in correspondence to 6V's as depicted within figure 1.

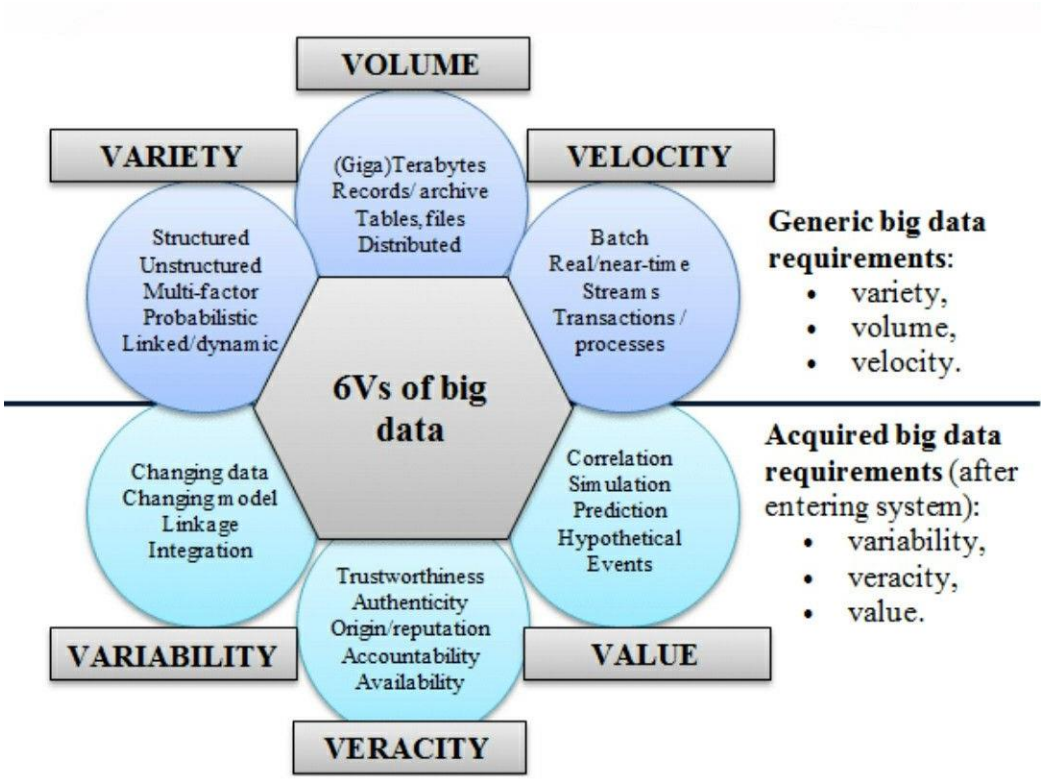

Figure 1: 6V's of Big Data (Safaa Alkatheri, 2019)

2.1.2 Big Data Analytics

A major aspect of big data is held within the analysis, as many academic scholars extend their definition of the term with the notion of analytical procedures, tools and techniques as stated by Bharadwaj et al. (Bharadwaj A, 2013). This definition is clearly perceived by Gantz and Reinsel, who have stated the Big Data Analytics (BDA) revolves around three main characteristics, data, analytics applied and the application of results in a means to generate business value (Gantz J, 2012). There are three specific steps required which depict how big data works, these steps include the collection of data, analysis of data and the application of knowledge (Ruth Dmonte, 2017).

Collection

- Collection corresponds to the first step and requires concrete technology. As immense data is generated it is critical to have specific technology, known as the Internet of things, to collect and store the data.

Analysis

- Once data is collected and properly stored, it requires proper analyses to reap the benefits of big data. The process of analysis facilitates superior business decision making, in hopes of improving performance. Commonly, data mining and data visualization are utilized for analysis

Application

- Once the analysis has been successfully outlined, the result can be used for predictions to grant a deeper understanding of other data.

In principle, "BDA encompasses not only the entity upon which analysis is performed -the data- but also elements of tools, infrastructure, and means of visualizing and presenting insight" (Patrick Mikalef, 2017). This notion of BDA has been further encompassed and supported by several other academics, as such concrete definitions are seen within table 2.

Author	**Definition**
(Kwon O, 2014)	Big data analytics: technologies (e.g. database and data mining tools) and techniques (e.g. analytical methods) that a company can employ to analyze

	large-scale, complex data for various applications intended to augment firm performance in various dimensions
(Lamba HS, 2015)	Big data analytics is defined as the application of multiple analytic methods that address the diversity of big data to provide actionable descriptive, predictive, and prescriptive results
(Müller O, 2016)	Big data analytics: the statistical modeling of large, diverse, and dynamic datasets of user-generated content and digital traces

Table 2: BDA definitions

2.2 Evolution of Data mining

The term data mining holds several unique disciplines, within the last decades these have been drawn to include five key aspects, depicted as database technology, statistics, machine learning, visualization and information science (Sumiran, 2018). The most prevalent use case of data mining techniques has been seen within the field of business, which depicts academic research dating to the early 1990s. Agrawal et al. concisely highlight the use of data mining techniques to predict employee actions or for external use to identify potential customers and or for the effective marketing of products (Agrawal, 1993). In line with the development of IoT, data mining has seen continuous use cases throughout all industries, as business-minded individuals understand the power which effective data mining holds. Since its offset data mining has developed and has generally been categorized according to two differential techniques: predictive and descriptive.

2.2.1 Predictive

Predictive techniques focus on the relational discovery amongst independent variables and its impact on a dependent variable. Predictive data mining has numerous utilities, these include forecasting explicit values dependent on patterns within the sourced data. Additionally, commonly includes the development of a model the likes of a neural network, to successfully predict a result. (Agrawal, 1993).

2.2.2 Descriptive

Descriptive techniques aim to describe a data set in a comprehensive fashion with aim of providing concrete characteristics of the data without any predefined target variable. There are several uses for descriptive data mining techniques, particularly within marketing which often aims to focus on intrinsic

structure, relations, the interconnectedness of the data. At its core, descriptive methods take data and highlight how things are related (Agrawal, 1993).

2.2.3 Data mining Techniques

The impact of the growing case of digitization, causing an exponential growth of data and hence big data analytics is an emerging trend and a dominant research field. A variety of techniques have developed over the course of the phenomenon, with concrete techniques playing an energetic role within a variety of industries. Algorithms provide an exposure to analyze, detect and predict (Majumdar, 2019). There are several different techniques, applicable to different use cases, techniques include Association, Classification, Clustering, Regression, Sequential Patterns, Visual data mining, Web mining, and other techniques (Sumiran, 2018). Concrete examples range from classification and clustering algorithm implementation for diabetes medical data (Majumdar, 2019), to Data Mining-based Data Replication (DMDR) in hopes of identifying the correlation between data files to improve replica management within the cloud environment (N. Mansouri, 2019).

2.3 Sports applications

The sporting world is known for its vast amounts of data, with each player having their concrete statistics collected, the likes of points, goals, assists and a variety of other data variables. To fully comprehend the usefulness of statistics within sports, it is first critical to analyze the software and hardware necessary to obtain the range of statistics, with the implementation of unique and groundbreaking IoT's within the realms of sporting competition.

2.3.1 IoT's within sports

Internet of Things holds a range of scholarly definitions, however, a recognized definition from Guillemin et al. states "the Internet of Things allows people and things to be connected Anytime, Anyplace, with Anything and Anyone, ideally using Any path/network and Any service" (P. Guillemin, 2009). IoT's are a relatively new frontier, however, have seen a vast array of uses and thus expanded exponentially, particularly within the fields of healthcare and wellness, sports and recreational activities, with these areas seeing a large growth in both personal and professional IoT technologies (Ray, 2015). With the growing use of IoT's and the surrounding research and development, numerous research articles dissect many aspects of interest and challenges faced when combining IoT and Sport (IoTSport). As such, two key frameworks have materialized delving into physical activity monitoring (PAMIoT) (Ray P. P., 2014) and Home Health hub Internet of Things (H3IoT) (Ray P. P., 2014). These frameworks highlight sensory inputs

received from the environment and how hardware processes data for it to be fed to the internet by applying network protocols thereby visualizing the comprehensive purposes. The real-world application of IoTSport not only sees recreational data visualization, with software applications such as Nike+ but also holds highly useful competitive applications with objective data and analytics through a network which can critically bolster performance. From a theoretical standpoint, the IoTSport framework utilizes a concept labeled ITPD (Interaction, Things, Processes and Data) as shown within figure 2.

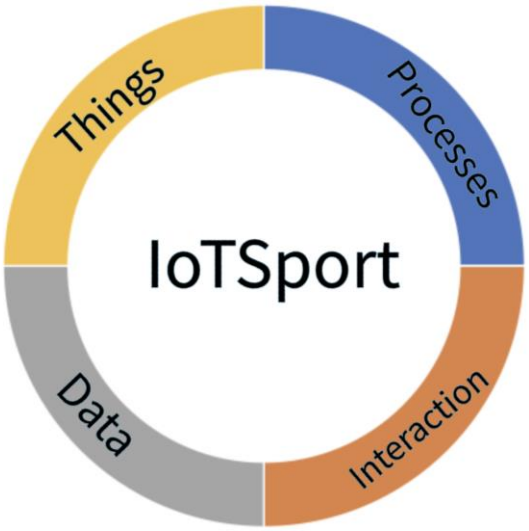

Figure 2: ITPD ring

Interaction:

- Involves athletes to get acquainted with sensors that may be present in their physical state or within their surroundings. Examples can include accelerometers, measuring everyone's speed of movement.

Things:

- Things involve the specific hardware, the likes of sensors, actuators, meters, and other measuring devices, which can be connected to the internet at any given period. Such hardware can be placed in any available environment, connecting to the network of distributed nodes and sharing information through the social web or personal cloud.

Processes:

- Processes include the business and/or technological actions needed to automate rapid growth in connections. The key tasks to be processed by IoT include Accumulation, Communication and Analysis.

Data

- There are several ways in which data can be collected, with the growing availabilities of hardware, data can be collected from sensors, chips, and other IoT connected. Data is collected in real-time, however, can be stored in the cloud or similarly analyzed in real-time. With the growing rate at which IoT is being developed, data can be streamed online to servers for immediate analysis, thus enabling competitive play to be altered within real-time.

In its entirety, the ITPD framework holds a powerful use when applied to the domain of sports.

2.3.1.1 IoTSport in professional sports

As noted, the IoTSport are not solely linked to recreational sporting behavior, however, play an increasing part within professional sports, in particular as associations invest more into understanding their players and how they operate through sensors, heartbeat monitors and other revolutionary IoTSport equipment. Not only is there a growing case for the use of IoTSport, however, there is a growing demand, which external companies the likes of Stat Sports are leveraging to develop. Stat Sports offers "player data to the most elite teams in the world" (StatSports, 2020), these include Arsenal, Juventus, the American Soccer team, Liverpool and others, leveraging their performance through the most powerful analysis software. To gather big data, there is a realm of connected devices, uniquely designed to gather specific individualized data. Within football, the use of IoTSport has particularly grown, as sensors have begun being implemented not solely during active gameplay, moreover during training sessions. The European Commission report on IoT in sports lists several devices actively used, these include ViperPod shirts, BodyCap capsules, ShockBox Helmets, and other devices designed to attain player data to increase performance (EuropeanCommision, 2017).

2.3.1.2 Challenges

Though the world of IoTSport has seen a rise from the offset of the 21st century, there remain a series of challenges in conjugation to IoT, largely due to the immature present state. Several challenges have been discussed academically and remain to be discussed today. Namely, Dmonte et al. have named three major challenges with regards to big data generated from IoT in sports, these include Data security, data transfer and lack of talent.

Data Security

- “It is essential to preserve the confidentiality of data” (Ruth Dmonte, 2017).

Data Transfer

- Generated data during sporting activities often yield data in a unique unstructured form, with sensors tracking player movement in real-time. However, it is often the case that organizations are unable to transfer the data in real-time.

Lack of Talent

- With the prevalence of new data forms in the world of sports, organizations are finding it difficult to “find data scientists, analysts who are capable of working on new technology” (Ruth Dmonte, 2017).

2.3.2 Big Data in Sports

With the continuous development of new IoTSport, sports are generating large amounts of data in relation to individual players, team performance and even fan interaction. A large variety of sporting events hold a high demand and fan interaction is often determined by performance, big data within sports and its need for deep analyses is seen as a must for future sporting associations. The success of big data, long before its inevitable use, has been documented within Michael Lewis’s book “Moneyball”, which as previously stated saw the rise of the Oakland A’s run by Billy Beane (Steinberg, 2015). Within the modern age of sports, the most prevalent use of big data is seen within the world of football, particularly seen by the German Football Association, who credited their 2014 world cup win to Big data technology. Similarly, the National Basketball Association (NBA) has seen a dramatic rise in its use of big data. These success stories have been denoted to use data technology known as SAP High-Performance Analytical Appliance (HANA), which is a form of BDA.

2.3.3 Big Data Applications in Sports

There are several unique use applications for big data within sports, several notions of big data have been discussed already, however within the world of sports there are concrete use cases which deem useful.

Match Results

- According to the founder of Advanced NFL stats Brian Burke, big data is very useful in determining match results, whereby big data is used to evaluate coaches and players to predict the results of a match and associations can take action accordingly (MIT, 2016).

Real-time statistics

- Big data offers real-time interaction, with large amounts of information available to fans through the likes of IBM's Slam Tracker which offers real-time analysis. An example includes point-by-point analytics during a tennis game (BDMS, 2014).

Team performance

- Big data is starting to be used, to grasp a deeper understanding of team performance. As organizations are becoming aware of the power that data holds, they can gain insight into their players to analyze their strengths and weaknesses, as well as their competitors, in hopes of improving team performance.

3.3.4 SAP HANA overview

SAP HANA is a relational database management system developed by SAP SE which operates with real-time data (Ruth Dmonte, 2017). The key role of SAP HANA is to provide a main-memory-centric data management platform that allows full support of SQL for traditional applications, additionally allowing for further expressive interactive models. The design of SAP HANA facilitates the handling of transactions and queries simultaneously; queries are typically written utilizing C++ or R programming. To fully understand how SAP HANA functions, one must understand the architecture behind the full relational database; SAP HANA architecture is depicted in visual form in figure 3.

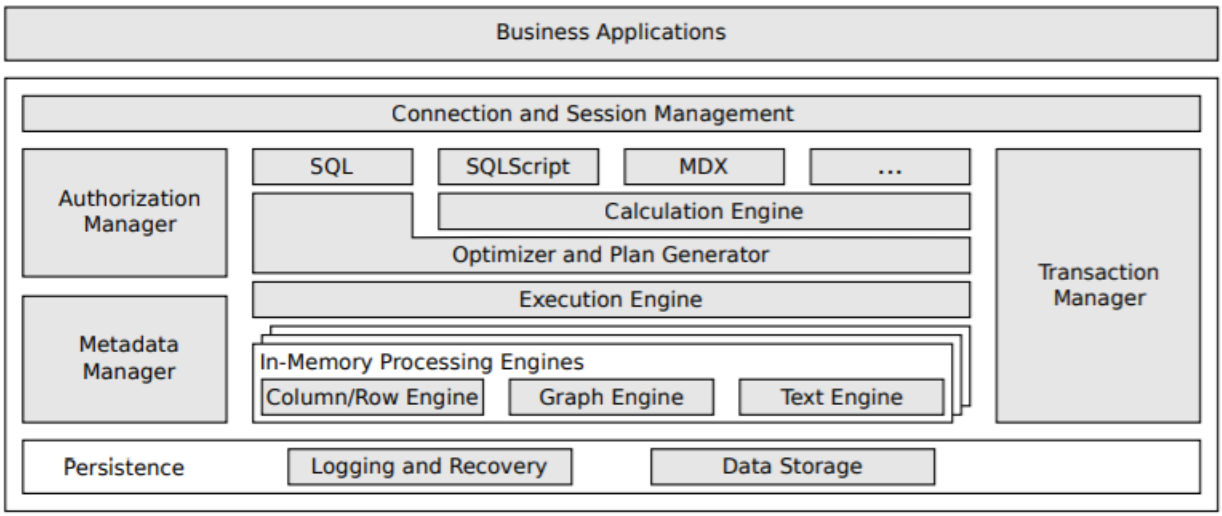

Figure 3: SAP HANA Architecture

The SAP HANA system design was established to be effective in allowing full transactional behavior to facilitate a wide array of interactive business applications, with concrete emphasis on parallelization ranging from thread to core level to a highly distributed system over several machines (Stratos Idreos, 2012). The built-in-memory processing engine is the major construct of the overall architecture, as it is the component that allows for a variety of data types to be stored. Enabled by the column/row engine where the relational data exists; the graph engine where graph data is stored, and the text engine which stores text data.

The further aspects of the SAP HANA architecture translate queries written in SQL, SQLScript and MDX. Particularly, the plan generator translates queries through the execution plan which is executed through the execution engine. Queries that are written using different programming languages are first described using the calculation engine (Ruth Dmonte, 2017).

Following the plan generator, the session manager takes the role of the middle tier, by connecting the front-end with the database, however first each user is passed through the authorization manager to check for the authorized user. Additionally, the transaction manager implements snapshot isolation to guarantee that all reads made in a transaction are consistent with the database at hand. The metadata manager is used for storing data which depicts tables and other forms of data structures within the system. Lastly, to prevent any data loss, the persistence layer acts as a contingency plan whereby backup data is stored in case of a system failure or system crash (Franz Farber, 2015).

2.3.5 SAP HANA Applications

As listed above SAP HANA offers speedy real-time processing and data analytics, which was prevalent during Germany's world cup victory in 2014. The German Football Association partnered with SAP to utilize their latest technology, playing a server role in Germany's success (Bojanova, 2014). SAP HANA was directly utilized through an app named 'Match Insights', which provided German coaches insights into their performance as well as their opponents' performance, key statistics included individualized passes, kicks, speed and player movement. Additionally, IoTSport monitored data points in real-time, such as IoT's included on-sight cameras and camera sensors. The real conversion from data to knowledge was seen with the application of SAP HANA, as data was transferred into the relational database, and provided coaches with detailed player performance analysis. By utilizing a software application, the SAP team one app was particularly useful in transferring real-time images, audios, or videos throughout the German Football Association. The usefulness of SAP HANA facilitated the storage of big data in both structured and unstructured form and was a major factor in aiding Germany's successful world cup run. The then-revolutionary use of big data was recorded within the wall street journal deeming it as "Germany's 12th Man" (Norton, 2014), providing an image of the real-time SAP interface, noted within figure 4.

Figure 4: SAP Germany's national team Snippet

Not only did SAP HANA play a major factor within Germany's world cup win, but SAP HANA is also playing a major active role within the American National Basketball League (NBA), with its presence being seen within the official NBA website 'NBA.com'. With over 118,000 minutes of play in an NBA season, the NBA generates a vast array of data ranging from player points, to distance ran, all of which holds valuable knowledge if properly analyzed (SAP, 2016). To properly utilize the data, NBA partnered with SAP HANA to provide an in-depth analysis of 50 years' worth of data statistics. Both structured data, in the form of

column/row data (points, assists, rebounds), and unstructured data, in the form of video, are inputted through SAP HANA to attain thorough game knowledge. One of the main reasons behind the use of SAP HANA within the NBA was to increase fan interactivity, and by developing the 'NBA experience' through the use of statistics the NBA has successfully developed fan engagement by 65% and seen upwards of a billion people access the NBA experience (SAP, 2016). Figure 5 highlights a visual representation of the NBA experience.

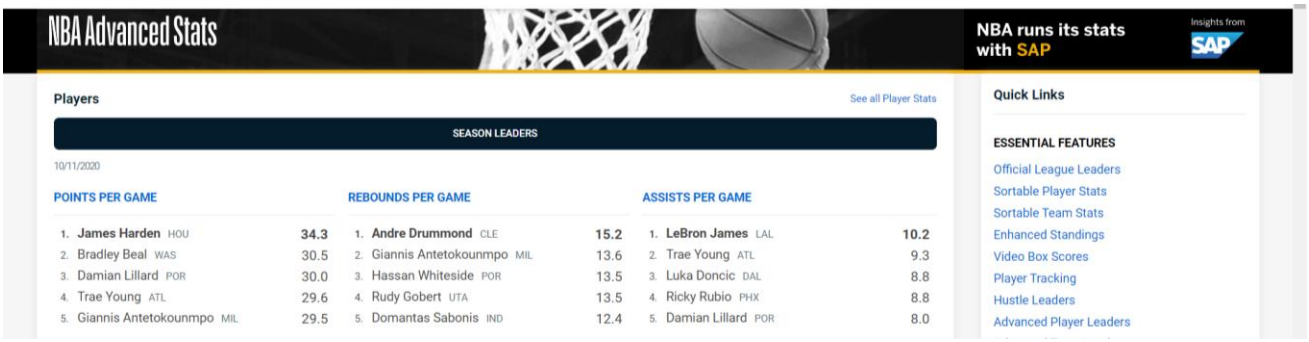

Figure 5: NBA Experience: highlighting the use of SAP HANA (NBA, 2020)

2.4 Data Mining in Sports

Throughout this chapter, the use of BDA in sports has been discussed, with the presence of the SAP HANA architecture, utilized to generate a wide array of contextualized knowledge. However, to further elaborate the topic at hand, data mining as a tool for analysis needs to be critically analyzed, hence within the following segment, both descriptive and predictive measures will further be expanded upon within the realms of sports.

2.4.1 Descriptive Data mining in sports

As discussed, descriptive data mining techniques aim to describe a data set in a comprehensive fashion with aim of providing concrete characteristics of the data without any predefined target variable. Within elite sports descriptive data mining ranges, with several descriptive techniques being used to extract performance patterns from previously held competitions. In particular, in 1997 Pincivero and Bompa recognized that to develop a deeper understanding of "the physiological systems within the sport of football are necessary in order to develop optimal training programs geared specifically for preparation as well as the requirements of individual field positions" (Pincivero, 1997). The two understood that for optimal training and therefore progress players needed to be analyzed on specific categories, namely body composition, strength, speed, and endurance. Techniques derived by Pincivero and Bompa are

recognized as one of the earliest forms of descriptive techniques within elite sports, since then specific techniques namely clustering, and classification have seen a growth rate of use. Within this project, clustering as a technique will be discussed.

2.4.1.1 Clustering in elite sports

Clustering is a means of unsupervised learning, utilized in finding how data is organized and summarizing/explaining key features of the data (Clausen, 2012). One of the most widely utilized methods of clustering within elite sports is K-means clustering; "K-means clustering is a clustering technique which is used to find an optimal number of clusters (K) which relate to the data set in such a way that the distance between the centers and the data points is minimized" (Wagstaff, 2001). Through K-means clustering, concrete groups are developed based on a set of variables where objects are grouped together based on similarities and objects of different groups are dissimilar as possible (Zachary Shelly, 2020).

A study performed by Shelly et al. highlights how k-means clustering can be and is being utilized within elite sports. Working alongside professional strength and conditioning coaches, their results provide a detailed analysis of how many groups are viable, with a test range of 3 to 7 groups. Their findings highlight that K-means clustering can effectively be utilized in training group players, with their final results highlighting that the technique is highly effective, however determining the optimal number of clusters is largely dependent on the association's facilities and how many resources they have to offer. Within the case-study at hand, the optimal number of clusters was deemed 3, with groups being labeled 'bigs', 'big skills and 'skills' (Zachary Shelly, 2020). Additionally, all athletes who showed characteristics between two groups were mathematically grouped according to real-time data gathered from concrete IoTSport technology. The three clusters are seen in figure 6.

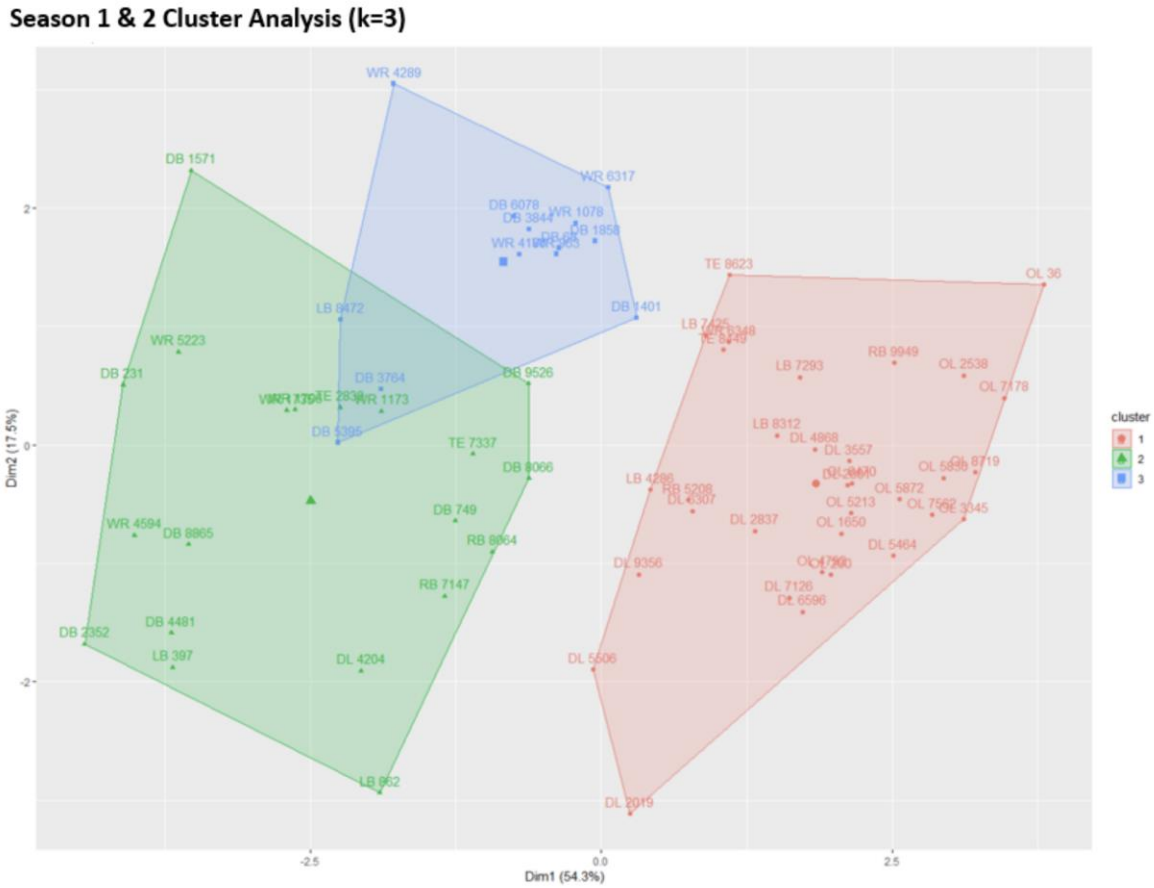

Figure 6: Group of clusters, generated from a student-athlete study (Zachary Shelly, 2020)

2.4.2 Predictive Data mining in Sports

Similarly, to descriptive data mining, within this chapter predictive data mining has been discussed and has been noted as the relational discovery amongst independent variables and its impact on a dependent variable. A highly discussed topic in athletics is the cost of injuries, as noted by Kerr et al. who identified that within the United Stated incurred injuries constitute a high proportion of healthcare costs (Kerr, 2015). The rising player values, as noted within the English premier league, which saw the average player value rise from £3.7million in the season of 2010/11 to over £12million in the season of 2019/20 (Sherlock, 2019), these figures are highlighted within figure 7, has seen the development of predictive models in aims of predicting future injuries. In order to facilitate effective model analysis, associations utilize concrete IoTSport, as noted within previous segments, to obtain wearable data which has been categorized into four main groups: physiological variables, mechanical load variables, locomotor variables, and acceleration, speed and distance variables (Caparrós, 2018). With the increasing evolution of IoTSport, associations have begun investing annually into new equipment, to obtain larger quantities

of data. This is especially the president when it comes to injuries, as the lack of data deemed a severe issue within the past, however, the "recent advances in wearable technology allow for real-time movement and load measurements for athletes during practice and training, with the availability of such versatile data, clinical prediction models are at the center of data-informed decision-making strategies to identify those individuals at high-risk for injury" (Amir Zadeh, 2020). Predictive modeling within sports, though rare, traditionally uses frequentist statistical models, or Bayesian models, as seen within Zadeh et al. study which saw a Bayesian approach to statistically determine the association between subjects and variables, to subjectively calculate the probability a subject (athlete) being injured with an associated confidence level. The study concluded that from 39 male candidates, who wore Zephyr helmets to facilitate data collection, candidates with two predictors had an injury incidence of 87.5% and cadets with fewer than two predictors had an injury risk of 39.1%, (Amir Zadeh, 2020).

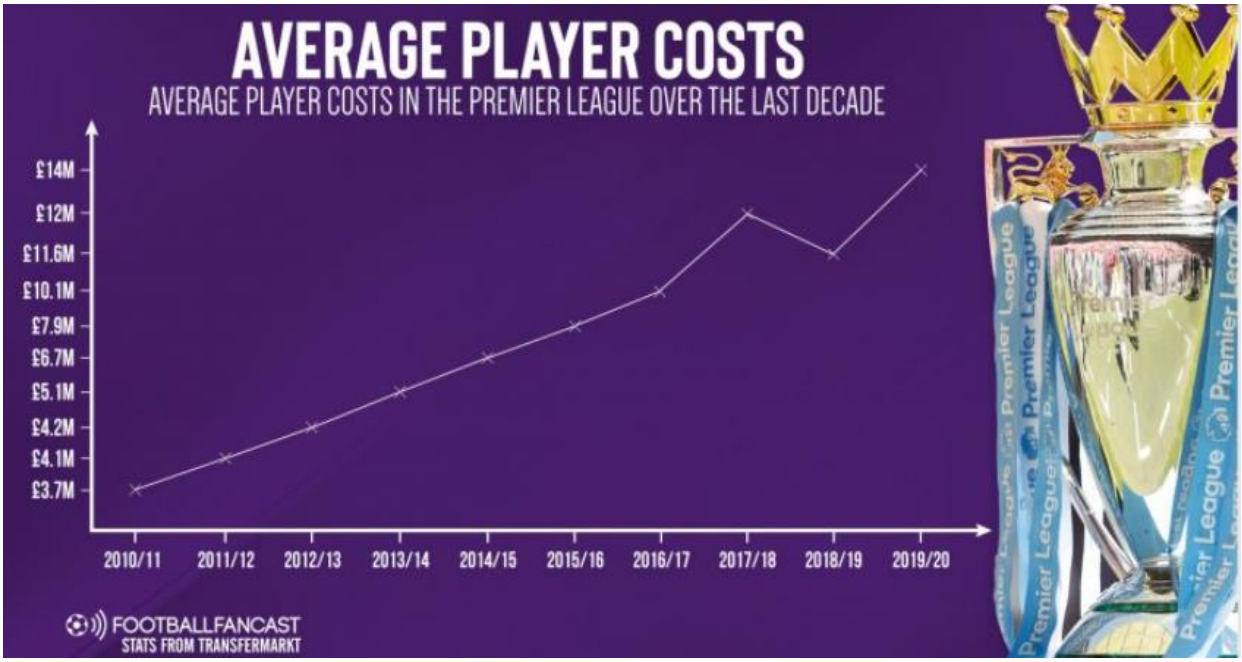

Figure 7: Average player values in the English Premier League 2010-2020

2.5 Introduction to the NHL

Thus far within the literature review, concrete topics have been discussed, from key definitions to the current uses of data mining within sports. Despite an overview of concrete studies and useful data mining techniques within elite sports, there remains a lack of concise data mining use when analyzing player salaries within the NHL.

2.5.1 National Hockey League

The National Hockey League (NHL) is the premier hockey league in the world. The NHL was established in 1917 in Montreal Canada and has grown to consist of 31 teams, spaced across North America; 24 teams are situated within the United States and the remaining 7 teams are situated within Canada. According to the official NHL rule book, teams are obliged to follow four different "roster" limits; a 20 player "dressed list" for games, a 23-player active NHL roster, a 50-contract maximum, and a 90-player maximum reserve list (Levin, 2009). Additionally, teams must abide by strict player salary regulations regarded as a salary cap. The salary cap comprises an annual set salary budget, set by NHL's board of directors, which each NHL team must abide by. The salary cap for the upcoming 2020/2021 season has been set to a "flat $81.5m" (O'Brien, 2020). There are many reported benefits to the enforced salary cap, as it has been regarded as the fairest system of play, as unlike other elite sporting leagues all teams are on a level financial playing field. Unlike in football, the NHL does not see vast disparities in roster values; within the English premier league, Manchester City's active roster comprises of a transfer fee equated to $1.1 billion, comparatively West Bromwich, a team within the same league, is estimated to have a transfer record of $110million (TransferMarket, 2020). Despite teams having a level playing field, the salary cap has seen teams squander large percentages of their budgets on 'playmakers', minimizing their remaining salary budget on remaining players.

The following segments within this Literature review chapter will focus on salary analysis, and how a predictive model could be used to illustrate future player values, to resolve a growing issue within the NHL, where teams are spending vast percentages of their salary budget on one or two players.

2.5.2 Salary Analysis

There have been several academic sports economic articles depicting the influence of salary caps in a professional team sports league. Staudohar gave a historical overview of the development of salary caps in American major leagues, identifying that the key reason behind the strategy was to entice equal opportunity amongst teams (Staudohar, 1998), despite some criticism, this notion has yet to change as other academics such as Dietl et al. hold similar claims, detailing the disparity amongst team revenues as a clear case for the salary cap (Dietl, 2009). Within the NHL the highest-earning team being the New York Rangers with a franchise value of $1650million, comparatively the Arizona Coyotes hold a mere franchise value of $300 million (Statista, 2019). The disparity between franchise values, merely supports academic claims, that a salary cap facilitates equal opportunity.

The current situation within the NHL has seen a rapid rise in extreme player salaries, with certain individual players holding an increasing percentage of a team's salary cap. Players such as Connor McDavid, Artemi Panarin, Auston Matthews and 10 others hold a cap hit of above $10million, which highlights the question, what makes a player worth millions?

2.5.3 Performance vs value

Published literature states that "a problem observed by NHL teams is that many players are being overvalued or undervalued for their performance, because their contracts, all other factors aside, are mainly monetized by one statistic: their total production of points (commonly referred to as plus/minus) for all games played" (Woo, 2018). Despite a player's plus/minus statistic being a very high factor in a player's performance, there are several variables that can be used to determine a player's performance, hence their value. Since, signing a multi-year contract with SAP in 2015, the NHL has begun to utilize a wide array of IoTSport, facilitating data growth. Since their partnership NHL has begun to publish an extensive presentation of enhanced statistics, namely shot attempts, shooting plus save percentage, alongside 30 new extended statistics (NHL, 2015). Hence following chapters will be answering how to predict the salary of an NHL player based upon performance data obtained from IoTSport technology.

2.6 Literature Review Summary

Within the literature review key factors surrounding the entirety of Sports data analysis have been discussed. The evolution of big data and data mining techniques highlighted the potential concrete methods hold, whether it be within sports or other industries. Proceeding this an overview of IoT's and their ability to obtain data through the ITPD framework provides a theoretical base, which was later applied to sports as the main topic of discussion. A major cohesive argument within the academic literature is the notion that through predictive and descriptive data mining is present within elite sports, certain topics have yet been properly examined. Player performance is a widely discussed topic, yet within the field of data mining, there are few academic papers that attempt to predict player values based on player performance, in hopes of identifying which players are overvalued or undervalued.

This research is distinctive for two primary reasons. First, many academic papers merely analyze generated big data from SAP, however, this project aims to go beyond and construct a predictive model utilizing available open-sourced NHL SAP data. Secondly, extends the existing frameworks for studying sports behavior, extending how associations should value their players, thereby determining their short- and long-term decision-making strategies.

2.6.1 Following the literature review

Within the remainder of this project, a predictive model will be constructed utilizing predictive data mining techniques. The aim of this model will be to mutually align player performance with player value and identify how associations can utilize this information to leverage decision-making.

3.0 Methodology

The methodology chapter will identify the central research methods which will be devised to tackle this project, with a key primary focus on research design. This section aims to identify the manner of achieving the approach to data analysis.

3.1 Research Questions

Within the introductory chapter of this project key research objectives for this project were formulated. Within the literature review chapter, concrete objectives were discussed utilizing academic journals and reports as key references to provide an in-depth background. The methodology chapter will rather focus on designing a predictive model, rather than providing insight behind the need for the model. Hence, this chapter will be split into two separate segments, one focusing on the design and theory behind the predictive model and the latter focusing on the application of the model third-party data.

3.2 Context

This project consists of an extensive literature review identifying key academic literature depicting key influential topics surrounding big data and the world of sports. Hence, to expand this research project and achieve listed methodology objectives, a Jupyter notebook has been constructed depicting a predictive model. “The Jupyter Notebook is an open-source web application that allows you to create and share documents that contain live code, equations, visualization and narrative text. Uses include data cleaning and transformation, numerical simulation, statistical modeling, data visualization, machine learning and much more” (Jupyter, 2020).

3.3 Design

Within the realms of data mining, there are various techniques that can be adopted to successfully develop a model. Within academic literature there are largely three popular data mining process models, these include Knowledge Discovery Databases (KDD) process model, CRISP-DM and SEMMA (Qaiser, 2014). For this proposal, the following will simply discuss the SEMMA model.

3.3.1 SEMMA

SEMMA is a data mining method developed by the SAS Institute and consists of five key steps, include Sample, Explore, Modify, Model and Access. Each stage will contain an in-depth analysis of the steps taken throughout the predictive model creation, however, below is a brief outline discussing the context of the problem at hand.

- Sample
    - The first stage of the SEMMA process focuses on the sampling of data. The NHL database consists of 2700 rows, and 3000 rows for the salary values, meaning that the data set will need to be sampled, to obtain a cohesive overall database.
- Explore
    - The second stage of SEMMA focuses on the exploration of data. This will be achieved through various libraries such as Pandas, which allow for graphical representations with the aim of discovering trends and possible anomalies.
- Modify
    - The third stage of SEMMA focuses on the modification of data, through the creation, selection and transformation of variables. In addition, the modified stage includes the reduction of the total number of variables (this will be achieved through the process of Principal Component Analysis) and the elimination of outliers.
- Model
    - The fourth stage of SEMMA focuses on modeling. A variety of models will be evaluated, the likes of KNN, ANN, and others, according to the concrete performance metrics a model of the optimal model will be fine-tuned.
- Assess
    - The final stage of SEMMA evaluates the reliability and usefulness of the findings and estimates the performance (Qaiser, 2014).

Figure 8 depicts a visual representation of the overall SEMMA phases.

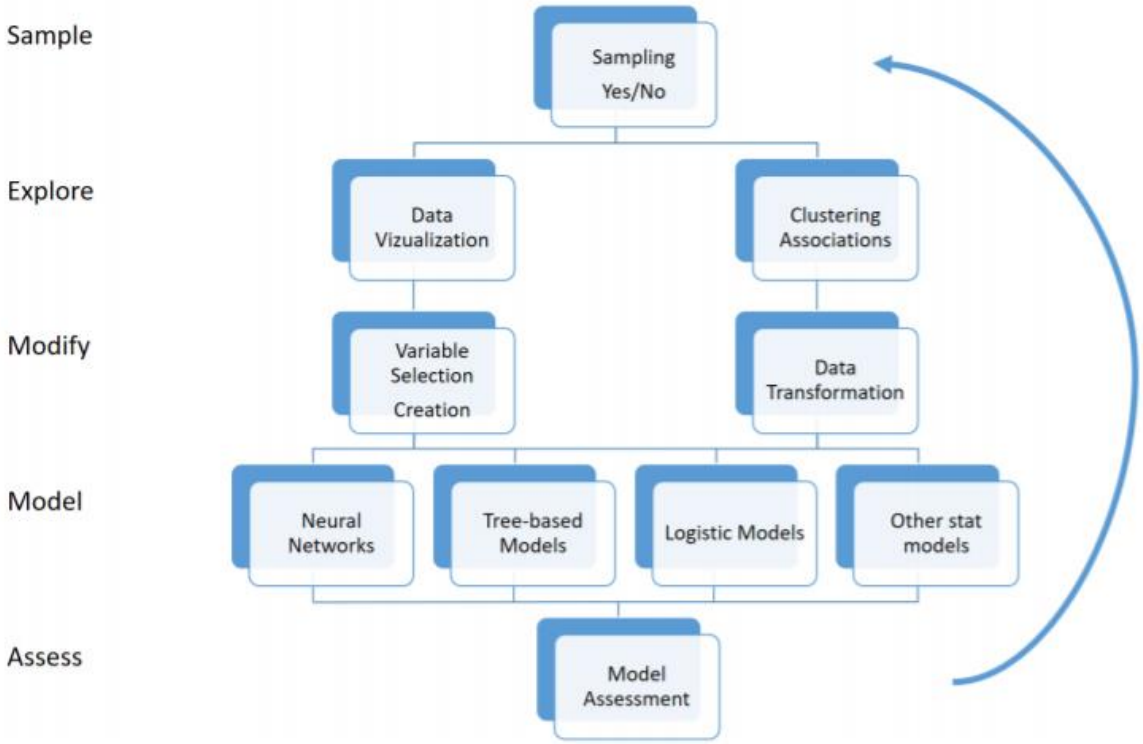

Figure 8: Different phases of SEMMA

3.4 Sample

The first phase of the SEMMA process, Sample, includes concrete stages of data collection and data partitioning into training, validation, and test samples. The following outlines the methods used for the context of the sampling phase.

3.4.1 Third-Party data

Third-party data has been used within the design of the predictive model. Open-sourced data consisted of historical NHL data, ranging from 2017-2007, consisting of 11000 rows and 24 independent variables. The salary dataset, presenting the dependent variable was separately contained, containing a total of 10000 rows and three variables. Since the two files contained the same composite key of player name and year, it was possible to concatenate them; the concatenated file consisted of 5920 rows.

To achieve a predictive model which could accurately predict player salaries a training, test and validation samples were needed. The test sample consists of player statistics from the year 2017, equating to 10% of the overall data. To produce a training and validation set, a 0.7/0.3 split was performed utilizing

train_test_split from the sklearn.model_selection library. A 70/30 split is a common practice whereby the model is built on the training set and applied to the validation set.

Initially, a model is fit on the training dataset, as an example to fit the parameters of the model. Once the model has successfully trained on the training dataset, the validation dataset is used as an unbiased model evaluation of a model fit on the dataset while tuning the model's hyperparameters. Lastly, the test dataset is used to provide an unbiased final evaluation by having the model fitted and exploring the difference between the predicted and actual target variable.

3.5 Explore

As part of the explore stage, the data will be 'explored' for any patterns and relationships, to gain an understanding of the data and draw any conclusive ideas prior to the modeling stage. Within this process two different techniques were utilized, for a quick and easy graphical scatterplot, Microsoft excel was used, and for a further in-depth exploratory analysis pandas profiling was utilized, as can be seen within the Jupyter notebook.

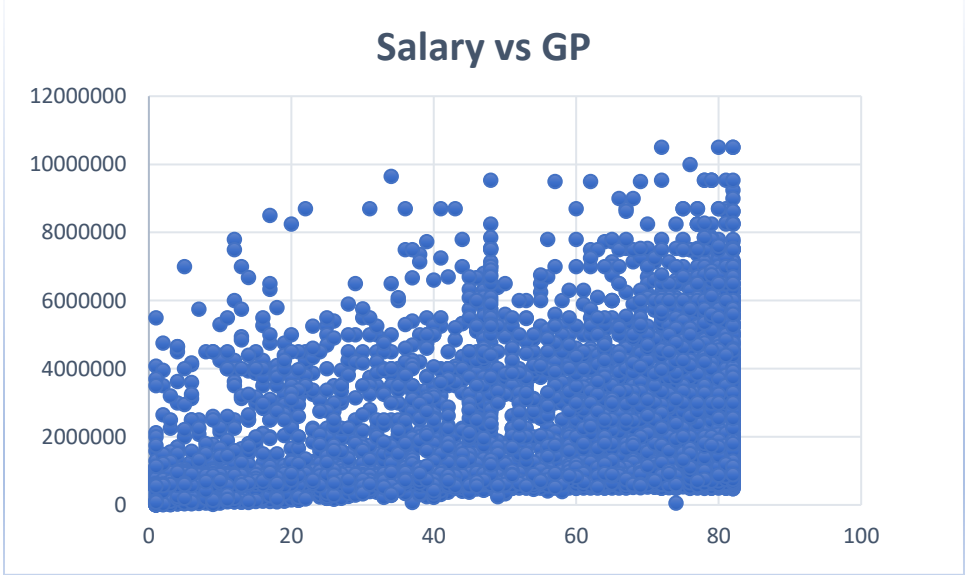

Figure 9: graph depicting how Salary is influenced by Games played

Although not perfectly accurate, it can be stated that players who are more active and have more games played tend to have an average higher salary.

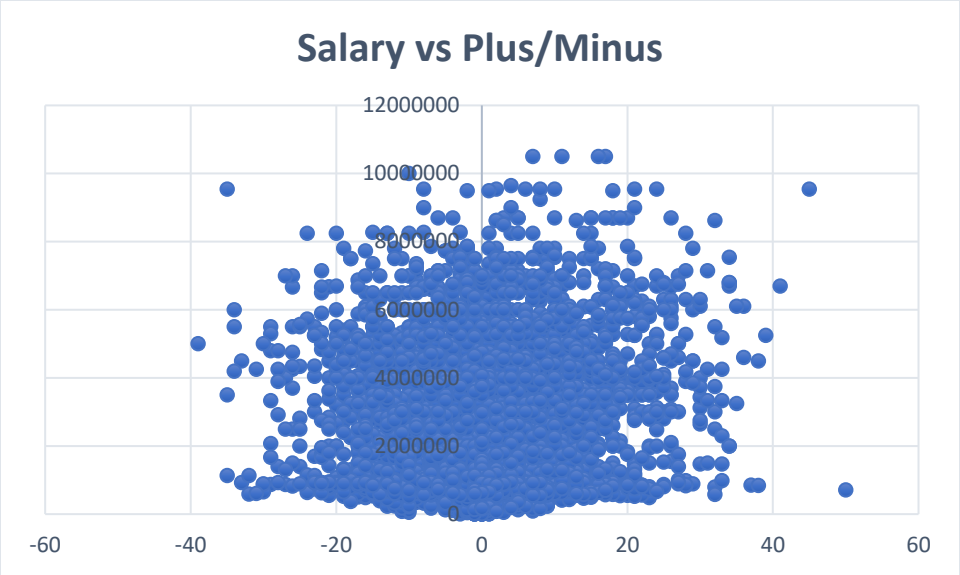

Figure 10: graph depicting how Salary is influenced by Plus/minus

Within the literature review, it was stated that common practice to measure salary was a player's plus/minus, yet the data concludes that many players hold a negative plus/minus yet hold a high salary.

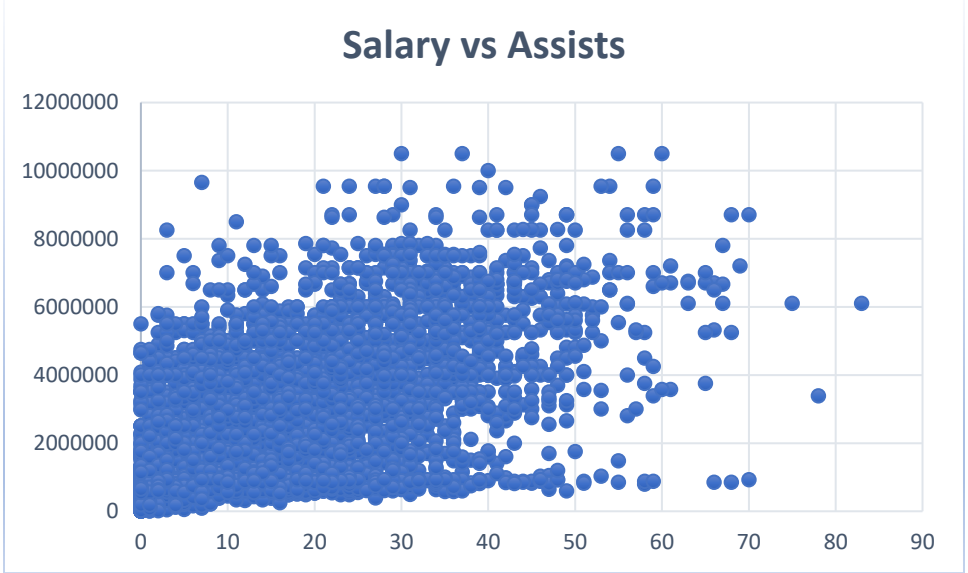

Figure 11: graph depicting how Salary is influenced by Assists gained

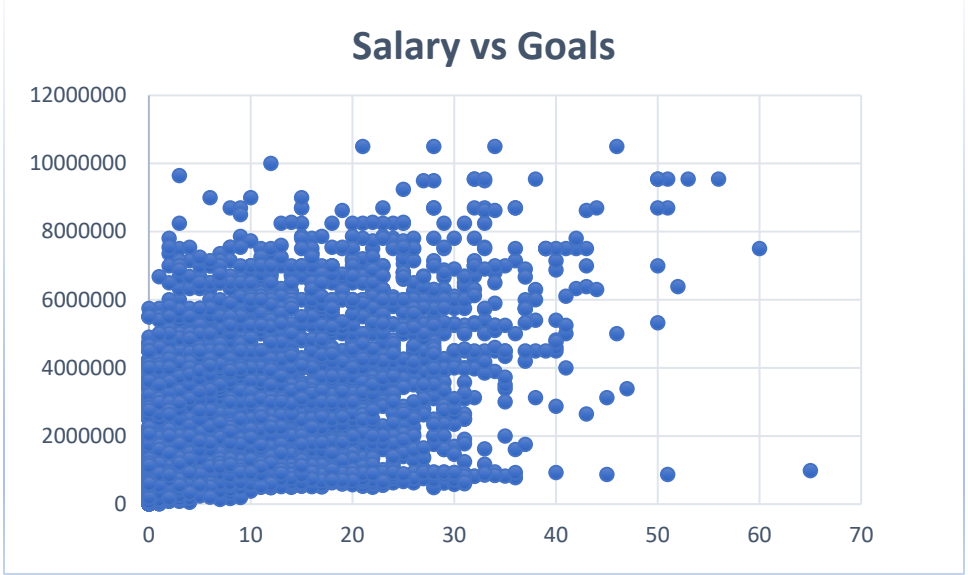

Figure 12: graph depicting how Salary is influenced by Goals scored

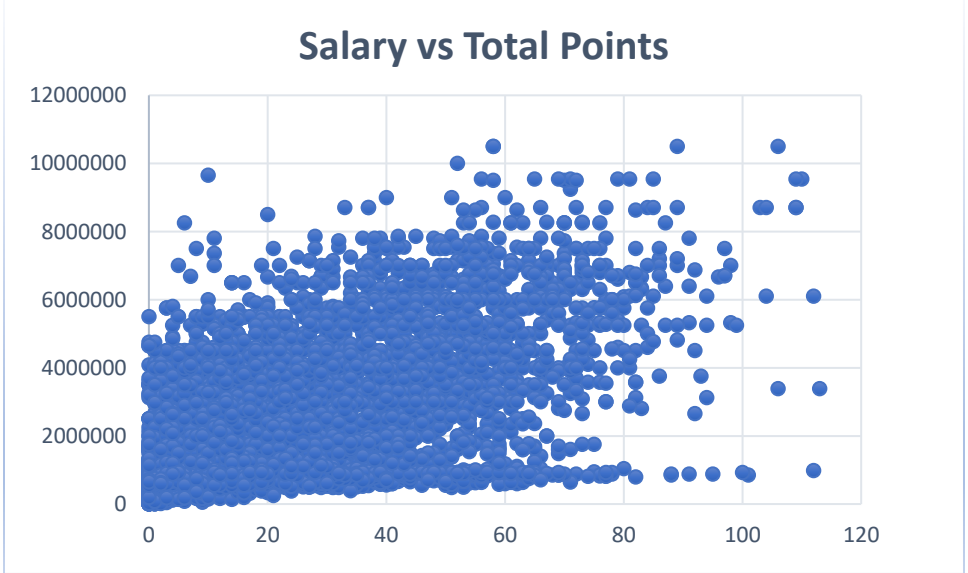

Figure 13: graph depicting how Salary is influenced by Goals scored

Figures 11, 12 and 13 highlights how salary is influenced by a player's ability to score and assist (set-up) goals, the common correlation is that the more assists/goals/total points a player obtains in a season the higher the salary.

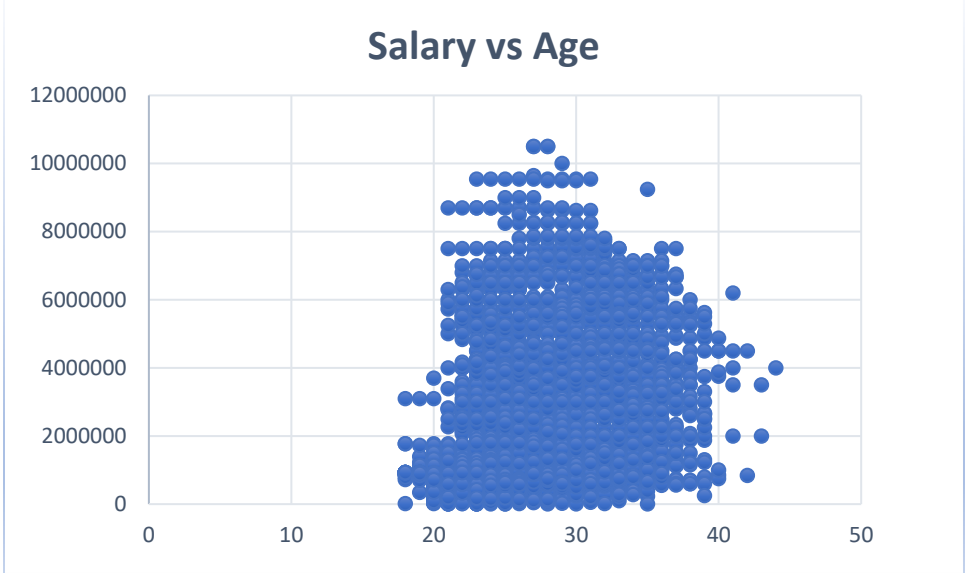

Figure 14: graph depicting how Salary is influenced by a player's age

It can be stated from figure 12 that younger players tend to have a lower salary, which may increase when they reach their 20' and 30's. However, it is clear that older players retain a lower salary, as most players within their 40's have a lower maximum salary compared to younger players.

Pandas profiling

The full pandas profile report can be seen within the Jupyter notebook; however, the major takeaways are presented below:

Overall

- There is a total of 16 numerical variables
- There is a total of 3 categorical variables
- 6 variables have been rejected due to being too highly correlated with other variables these include:
    - Points were rejected due to having a high correlation score of 0.959 with Assists
    - Point shares were rejected due to having a high correlation score of 0.91 with Total points
    - Even Strength Goals was rejected due to having a high correlation score of 0.966 with Goals scored
    - Shots were rejected due to having a high correlation score of 0.91 with Total points
    - Time on Ice was rejected due to having a high correlation score of 0.91 with Games played

- Face of Losses was rejected due to having a high correlation score of 0.98 with Face of Wins

Salary

- The mean salary of an NHL player is $2.2million, with the maximum value present within the database being $10.5million.
- The lowest salary is $2957, this value is an anomaly as the lowest cap salary offered to an NHL player cannot be lower than $450 thousand (NHL, 2020). This anomaly is dealt with within the feature engineering section of the Jupyter Notebook.

Age

- The mean age of an NHL player is 26.63 years of age.
- The youngest players within the NHL are 18 years of age, which the minimum requirement when joining the national hockey league.
- The oldest player within the database is 44 years of age, this player being Jaromir Jagr who was regarded as the 'The Ageless Wonder'.

Goals/Assists

- The highest recorded number of goals was 65 by Alexander Ovechkin
- The highest recorded number of assists was 83 by Henrik Sedin
- The mean number of goals for an NHL player is 8.3 goals per season; the mean number of assists for an NHL player is 14.17.

Plus/Minus

- Within this database, the mean plus-minus for an NHL player is 0.155.
- The highest recorded plus-minus for a player in a season was 50 by Jeff Schultz
- The lowest recorded plus-minus for a player in a season was -39 by Alexander Edler

3.6 Modify

The steps within the modification segment of SEMMA builds upon the previous Explore step. Within this step, data begins to be modified and prepared to be used for a specific model (Antonia Daderman, 2018). The process of modifying the NHL dataset consisted of data pre-processing. The steps taken within data pre-processing are discussed below.

3.6.1 Feature engineering

"Feature engineering is a crucial step in the process of predictive modeling. It involves the transformation of given feature space, typically using mathematical functions, with the objective of reducing the modeling error for a given target. However, there is no well-defined basis for performing effective feature engineering. It involves domain knowledge, intuition, and most of all, a lengthy process of trial and error" (Udayan Khurana, 2017). In accordance with the definition, certain variables contained within the NHL database were modified and new variables were created and added.

Average Time on Ice

Prior to the deletion of Time on Ice, a new variable was created depicting the average time a player spends on ice per game. To calculate the average time on ice, the time on ice was divided by the games played.

Player ID

Since the database consists of historical NHL data through the period 2007-17, certain players would repeat annually however hold different statistical values. To separate each dataset, a composite key was created by combining a player's name and the year for which the dataset was represented.

Salary

For the sake of visual simplicity, the salary target variable has been divided by a factor of 1000, thus rather than having the entire salary value, the data frame holds player salaries in terms of 1000's.

Dummy Variables

Dummy variables are needed to convert categorical variables to numerical values. Particularly within regression analysis, dummy variables are utilized to convert categorical values to only take the value of 0 or 1 (Gujarati, 2003). Within this project, dummy variables were used to convert the Team variable and a player's position variable.

Salary Anomaly

Within the profiling, it was noticeable that some players had salaries below the NHL threshold, hence players who had a salary value of below $450 thousand were dropped from the data frame.

Deletion of highly correlated variables

The process of deleting correlated variables only occurs when the correlation is so strong that they do not convey any extra information. To understand the correlation amongst variables a secondary pandas profilin report was conducted. The conclusion reported that a total of five variables yielded correlation scores which portrayed no further information, hence were deleted. These variables included: Total points, Face of losses, point shares, even-strength goals and Shots taken.

3.6.2 Missing Data and Duplicates

Missing data or incomplete data can lead to poor data analysis alongside wrong analysis results. Furthermore, the presence of missing data or duplicates leads to the destruction of the 'Veracity' of big data (Wang & Jones, 2018). There are various techniques available when dealing with missing values and duplicates, these are identified below.

Duplicates

Fortunately, within the NHL database, there were no present duplicate datasets. This fact is seen within the Jupyter notebook.

Missing Values

There are countless techniques for the handling of missing data. Anesthesiol et al. state the most common practice when dealing with missing data is to omit the missing data and analyze the remaining data. Additionally, they state that other techniques include listwise deletion, pairwise deletion, or mean evaluation, stating that the main fact is the volume of missing data and the preference of the data analyst (J Anesthesiol, 2013).

Within this project, the technique used was based on popular academic practice, the removal of missing values. Missing values were presented within three variables: Shooting percentage, Blocked shots, and Faceoff percentage. Shooting Percentage had 1.37% of missing values; blocked shots had 0.0175% of

missing values, and face-off percentage had 34.28% of missing values. Since face-off percentage saw such a large proportion of values missing, the variable was omitted from the data frame.

3.6.3 Data Normalization

Normalization is a critical stage in all decision models, with its ability to produce comparable and dimensionless data from heterogeneous data. The process of data normalization is available in a variety of techniques, with performance depending on a series of characteristics of the problem at hand (Nazanin Vafaei, 2019). Various techniques of normalization include Z-score normalization, Normalization by decimal scaling, min-max normalization and other forms. Within this project, normalization has been utilized in attempts of giving all attributes equal weights. In particular, the means of normalization used is known as Min-max normalization, which performs a linear transformation thereby preserving the relationships among the original data values (Jiawei Han, 2012). Within the context of this project, min-max normalization was used due to its simplicity alongside the fact that the target variable (Salary) highlights a large disparity between the smallest value and the largest, min-max normalization offers the best feasible solution to compacting the data within a 0-1 scale.

3.3.4 Data Splitting

Within machine learning, a key requirement for computational modeling is the process of splitting. Within this project, the use of simple random sampling was used, as it is efficient, easy to implement and the most common practice (Reitermanov´a, 2010). Simple random sampling splits data into a randomly selected sample with a uniform distribution. There are several advantages why this approach was used, key being that it leads to a low bias of model performance and the fact that each sample has an equal probability of selection. The ratio between samples was chosen to be 70% for training and 30% for testing on accord with academic papers. Reyes et al. stat stated that it is common to separate data in the ratio of 7:3 for training and testing respectively (Jeremiah Reyes, 2015).

3.7 Model

The model stage consists of modeling the NHL data through concrete software which facilitates a combination of data that reliably predicts the target salary. The sklearn model selection library was used to identify key regressor models, with a total of six models being chosen. To successfully evaluate each model, functions were developed to generate overall metrics for each model, in attempts of comparing each constructed model.

3.7.1 Regressors & Evaluation

There are several different cases of regressor models, the sklearn model selection library offers a wide array of unique models each with its own advantages. To identify which concrete regressor model was deemed best fit within this project, specific evaluation metrics were identified, to numerically evaluate each model, thus facilitating a comparison amongst key models. Chosen regression models include the following (each model has been described within the Jupyter Notebook)

- Baseline Model, using DummyRegressor
- Linear Regression, using LinearRegression
- K-Nearest Neighbors model, using KNeighborsRegressor
- Decision Tree model, using DecisionTreeRegressor
- Random Forest model, using RandomForestRegressor
- Neural Network model, using MLPRegressor

Additionally, a set of four evaluation metrics were used to identify the optimal model. Evaluation metrics are listed below with their appropriate numerical meanings.

Mean Absolute Error

- Within statistics the mean absolute error represents the measure of errors between paired observations. Mathematically the MAE is calculated utilizing the following equation. Thereby the MAE score is the arithmetic advantage of the absolute errors.
- $MAE = \frac{\sum_{i=1}^{n} |yi - xi|}{n} = \frac{\sum_{i=1}^{n} |ei|}{n}$
- Since the MAE calculates the absolute error represented within a model, the aim is to minimize the MAE as far as possible.

R^2 Score

- Within statistics, the R^2 score represents the proportion of variance in the dependent variable (Salary), with respect to independent variables. The primary purpose for an R^2 score is notably within prediction as it provides a measure of how well-observed outcomes are replicable by a model, in accordance with variation presented by the model. Mathematically the R^2 score is calculated utilizing the following equation.
- $R^2 = 1 - \frac{Unexplained\ Variation}{Total\ Variation}$

- R^2 score ranges between the values of 0-1, with higher values representing a higher linear association (Condor, 2020)

Adjusted R^2 Score

- Within statistics, the adjusted R^2 score is utilized to compare the goodness-of-fit for regression models which contain different numbers of independent variables (Frost, 2020). The adjusted R^2 is dependent on whether a new term improves the model's goodness-of-fit. Mathematically the adjusted R^2 score is calculated utilizing the following equation.
- $R^{-2} = 1 - (1 - R^2)\frac{n-1}{n-p-1}$
- Contrary to the R^2 score, the adjusted R^2 score is measured on a scale from -1 to 1

Symmetric mean absolute percentage error (sMape)

- Within statistics, the sMape is a measure of accuracy based on the percentage of errors. Mathematically the sMape is calculated utilizing the following equation.
- $sMAPE = \frac{100\%}{n}\sum_{t=1}^{n}\frac{|Ft-At|}{(|At|+|Ft|)/2}$
- sMape has a range from 0-100, with a smaller value indicating a lower percentage of error.

Listed evaluation metrics were utilized to identify which model optimized the goodness-of-fit, based on the NHL dataset. The represented generated values are highlighted in figure 15 below.

Summarized Results

Train

Model	MAE	MSE	R^2	Adj_R2	sMAPE
Baseline Model	1645	3895718	0.0000	-0.0122	75.5
Linear Regression	970.4	1657182	0.5746	0.5624	54.16
KNN	823.7	1352361	0.6529	0.6486	39.1
Decision Tree	0.1652	6.83	1	1	0.0227
Random Forrest	305.1	199665.7	0.95	0.9481	16.2
Neural Networks	770.65	1102147.9	0.7171	0.7136	43.15

Test

Model	MAE	MSE	R^2	Adj_R2	sMAPE
Baseline Model	1617.5	3739622.5	0	-0.029	74.4
Linear Regression	957	1636281	0.5624	0.5498	52.5
KNN	1058	2125552	0.4316	0.4152	49.3
Decision Tree	1084.64	27977380.8	0.252	0.2303	45.1
Random Forrest	806.4	1364337	0.6352	0.6246	37.35
Neural Networks	884.1	1435550	0.6161	0.6050	48.6

Figure 15: Summarized results, split according to train and test values

In accordance with the definitions of utilized metrics, the optimal model was identified as the Neural network model. Even though the neural network model does not have the lowest MAE, MSE, sMAPE, neither does it have the highest R^2 and adjusted R^2 score, it yields the lowest divergence between the train and test results. Contrary to the Random Forest model, which portrays the best train metric scores, the Neural Network model does not portray a similar case of overfitting; hence, for further model optimization, the Neural Network model was chosen.

3.7.2 Model Optimization

Model tuning is the process of optimal parameter selection and is an essential aspect of numerical modeling. Even though within the segment above, the optimal model is seen to be a Neural Network model, further model optimization is necessary, to fine-tune model hyperparameters. “The process of estimating these uncertain parameters in order to reduce the mismatch between specific observations and model results is usually referred to as tuning” (Hourdin, et al., 2017). Within this project, to correctly identify the optimal MLPRegressor hyperparameters a process of random search tuning alongside grid search tuning was performed.

3.7.2.1 Random Search

The notion of a random search is the use of sample points as a means to move toward a global minimum point, the method utilizes the "idea of simplex, which is a geometric figure formed by a set of n+1 points in the n-dimensional space" (Arora, 2004). Academic literature states that a less complex optimization technique, the likes of Random Search (RS), may be sufficient for hyper-parameter optimization (Rafael G. Mantovani, 2015). Bergstra et al. have claimed through their extensive research study that though RS is a simple technique for model optimization, empirical and theoretical evidence highlights that RS is a highly efficient means for hyper-parameter optimization. Furthermore, it has been stated that compared with neural networks configured by a pure grid search, studies show that RS over the same domain is able to find models which are as good or better within a small fraction of computation time (Bengio, 2012). Henceforth, as the first stage of model optimization, RS was utilized to generate a superior model than the base MLPRegressor model. The selected hyper-parameters within the random grid included hidden_layer_sizes, activation function, alpha, learning_rate and the maximum number of iterations.

In accordance with RS, a further method of tuning (Grid Search) was utilized to further narrow the MLPRegressor hyper-parameters.

3.4.2.2 Grid Search

According to Bengio et al., Grid Search (GS) is one of the most widely used strategies for hyper-parameter optimization (Bengio, 2012). The process of GS involves "taking n equally spaced points in each interval of the form [a,b] including a and b, this creates a total of n^m possible points to check, finally, once each pair of points is calculated, the maximum of these values is chosen" (Jean-Marie Dufour, 2019). Typically, GS is a method that is deemed un-efficient since with an increasing number of parameters computation time vastly increases. As the number of parameters increases, the number of evaluations increases exponentially, henceforth RS was performed prior to GS to obtain a rough notion of an optimized MLPRegressor model, and merely optimize it further by running GS on merely three hyper-parameters. Within this project, the chosen hyperparameters were the hidden_layer_sizes, alpha and the maximum number of iterations. By choosing merely three hyper-parameters the number of fits was set to a mere 9600, equating to a computational time of 309.4minutes.

3.7.2.3 Summary

As a result of both RS and GS, the MLPRegressor model, yielded superior evaluation metric scores. Figure 16 highlights the new metric scores attained for the constructed optimized model.

```
Train: MAE = 645.9071
Test: MAE = 770.9256.

Train: MSE = 866051.7989
Test: MSE = 1245054.7369.

Train: r2 = 0.7721
Test: r2 = 0.6586.

Train: r2_adj = 0.7693
Test: r2_adj = 0.6487.

Train: sMAPE = 30.1111
Test: sMAPE = 34.3012.
```

Figure 16: optimized MLPRegressor metrics

As highlighted within figure 16, the newly optimized model has scored superior on all fronts. Compared to the base model the MAE, MSE and sMape are lower for both train and test sets. Furthermore, both the R2 score and the adjusted R2 score have increased significantly for both train and test sets.

3.8 Assess

The assessment stage and final stage of the SEMMA process consists of assessing the data and evaluating the usefulness, reliability of the findings from the data mining process and estimate how well it performs (Santos, 2008). To fully understand the usefulness of model findings, the validation set comprising of 2017 NHL player data was fit, to fully evaluate the findings and how they can be used for effective decision making. The full assessment of constructed model is discussed within the next chapter.

4.0 Data Analysis

This chapter will delve into technical applications of the MLPRegressor model and its ability to evaluate player performance. Within the introductory chapter concrete research objectives were stated, hence within this chapter, the aim is to analyze the constructed data model and identify how effective use can facilitate superior decision making.

4.1 Model Interpretation

Once the constructed model was fitted to the validation test set, the model yields a sufficient means of predicting salaries. The excel spreadsheet depicting the predicted salaries in comparison to real player salaries is presented within the 'Salary_Pred.csv' file and highlights the actual values, alongside what the model predicted, a snippet of the CSV file is seen within table 3 in the Appendix, at the end of this project. As is portrayed within the CSV file, the predicted model has an average error of $150 thousand and a Mean Absolute error of $870,000, meaning the model can be used as an effective means of predicting player salaries, based on their seasonal performance statistics. Throughout the literature review, particular literature has been analyzed, with the underlining conclusion cementing the notion that though big data and the implementation of IoTSport is becoming the norm within professional sports, further data mining techniques can and should be utilized to support decision making. The constructed model, though highlighting error, within the context of the NHL and its use of the salary cap, general managers and associations would be able to attain knowledge for future contract negotiations. In particular, as depicted within most academic literature the MAE is a widely adopted recommendation system to measure the difference between predicted values and true absolute values. Thus, a MAE of $870,000 means that if NHL organizations were to use the constructed predictive model player salaries would lie within the MAE range. To fully take advantage of the knowledge gained from the predictive model, an ideal combination of tacit knowledge, held by coaches, managers and other personnel, alongside new constructed knowledge has the potential to successfully alter an organizations decision-making.

4.2 Player negotiations

Since the introduction of the salary cap, each team within the NHL has had equal opportunity to formulate contracts with players they deem fit for their organization, with certain teams governing their resources into a handful of players. This notion is seen within several NHL teams that invest highly into their starting lineup as opposed to the overall team. The Edmonton Oilers are a key example who hold five players with a salary equating to 44% of their total available salary cap, though this project does not cover if this is an

effective strategy, it does bode the question of whether teams should merely focus their resources on top prospects or spread them across the entire roster.

Negotiating is a key skill that is required within professional sports, with academic literature from the early 1990s depicting the art of negotiation as a crucial art form that can 'make or break' a competitive association. Falk listed in his book "Art of Contract Negotiation" that to be successful when negotiating player contracts, one must "be an expert on the salary cap and has to be creative and try to invent ways around the cap" (Falk, 1992). To succeed within an environment, which requires an underlying understanding of variables that are not at one's fingertips, necessitates knowledge. As noted within the introductory chapter knowledge is established from the conversion of raw data into information and information into knowledge. Concretely literature elaborates the power of knowledge and states "The value of knowledge derives from the intensity and the range of its use" noting it is critical to maximizing the amount of knowledge when developing a competitive advantage (Johan Muller, 2019). This notion of knowledge is power, is further accentuated within professional sports, as understanding how much a player is worth can alter a team's future competitive opportunities. Notable examples include the successful signing of Tyson Barrie, who was signed for a mere $3.75million cap and tallied an impressive 59 points. Contrary, one of the worst signings include goaltender Mike Smith who was subject to one of the worst seasons with a mere 0.902 save percentage attained a salary cap of $4.25millions (CapFriendly, 2020). Henceforth, the constructed model could have been utilized in both identifying which players to approach and how to operate throughout the negotiation process.

4.3 Model insights

The validation set consists of 769 players and holds vast information which general managers could utilize to construct knowledge; a clear case is the example of Connor McDavid. Connor McDavid is the captain of the Edmonton Oilers and is currently ranked the best player within the NHL. Based on the model during the 2016-17 season, though McDavid's contract salary cap was $925 thousand his predicted salary was $3.7million more at a value of $4.7million. This example merely highlights how it could be utilized to identify which players are either undervalued or hold a high potential, thus could exceed in future seasons. Additional cases based on the model include players the likes of Eric Staal who is according to the model undervalued by a figure of $3.4million, and Dylan Larkin who is undervalued by a figure of $3.5million, both players have since outperformed their 2016-17 season with Eric Stall recording 11 more points and Dylan Larking recording 31 more points per season since. Contrary, the model can depict which players are deemed overvalued or are past their prime, players such as Justin Abdelkader, Andrej Sekera

and Mark Methot. All three of these players based on their overall performance have contributed less annually, with each player contributing to fewer points and having a smaller plus-minus. As such, through the effective usage of the constructed model, associations would be able to depict which players have the potential to yield a superior performance than their current contracts. Henceforth, facilitating a deeper understanding of future player performance and offer contract salaries accordingly.

4.3.1 Model Limitations

Despite accurately predicting the growth of certain players, the constructed predictive model holds concrete limitations which have the potential to negatively impact an organization's team composition. The above segment merely lists players that were correctly evaluated and highlight how their player growth, since 2017, matches the predictions generated by the model, however, the same model has also generated predictions that are outright contrary to players' performance statistics. A prime example is seen with the player Filip Forsberg; Filip Forsberg is a key player for the Nashville Predators and has a player salary cap of $6million, yet the constructed model has his estimated worth at a significantly lower value of $1.9million, a difference of $4.1million. According to the constructed model, Filip Forsberg should not hold such a high value, and teams should not invest such significant resources, yet as a player he has merely grown and developed, scoring more goals, yielding a higher plus/minus and contributing to an increasing number of overall points each season. Filip Forsberg is merely an example of several players, who have highlighted continuous development, yet the predictive model estimates their worth at a lower value.

Within the world of elite sports, several factors determine star quality, within this project concrete numerical variables have been utilized to predict a player’s salary, however, there are additional factors that cannot be quantified that hold a significant role in determining a player's worth. Though there is limited academic literature, Santos et al. provide a descriptive evaluation of a positive teammate psychological development in high-performance sport, stating “team captains play an important role in promoting positive life-skills development in their teammates” (Fernando Santos, 2017). Within elite sports, the role of a leader is often a key factor in determining seasonal success. Within the NHL, the role of the captain is particularly vital; Dave Ungar has stated that captains are more important in hockey than any other major sport, as “in hockey, the team captain might not be close to being considered the best player, on the team”, but exhibits traits that can make the overall team a Stanley Cup contender (Ugnar, 2012). Specific players the likes of Jamie Benn depict these characteristics; Jamie Benn is the captain of the Dallas Stars and leads his team to the Stanley cup finals in 2020, however, according to the constructed

predictive model, his worth is substantially lower than his actual salary. Lack of academic literature has yet seen a direct means of quantifying personality traits, thus players which exhibit beneficial traits, however, do not highlight other numerical variables that may be undervalued by the constructed predictive model. This idea will be further examined within the following segment labeled 'further research'.

4.3.2 Impact on decision making

Ward et al. state the application of "scientific principles to inform practice has become increasingly common in professional sports, with increasing numbers of sports scientists operating in this area" (Patrick Ward, 2018). In addition, Ward et al. suggest that the fundamental process of data collection, data processing and data modeling assist in the development of robust decision-making processes throughout an organization. Through effective usage, data analysis ought to be utilized to determine decisions, through tracking outcomes and an effective feedback loop system a variety of decisions can be determined to be successful or pose a risk. Analytical models are a concrete area which if effectively utilized can alter the decision-making process; within professional sports, models have the power to determine outcomes prior to a pre-emptive decision. Within the NHL there are a handful of concrete decisions which have resulted in significant long-term results, from offering an excessive player contract to poor trades. Success within hockey is the process of wise decisions, and historically the notion stood that "it is the things that you do not know, which will cause the greatest pains in the game of hockey" (MacDonald, 2018), however with the introduction of new IOTSport and proper data analysis, organizations have the ability to correctly evaluate their strategy. A result of the constructed predictive model can be superior decision making, by utilizing the knowledge created, organizations have the ability to set salary limits on players they deemed fit, rather than risking resources on players who are projected to succeed according to word of mouth. By combining intuition and model analysis, teams could effectively determine the impact of their decision-making, thereby enabling them to construct a roster that has the potential to compete.

4.4 Summary

The continuous stride within information science, through major advances in hardware and software, is facilitating an endless array of data analysis, enabling the conversion of data to knowledge. Through data analytics, not only is information produced from data, but further actions and decisions build upon the growth converting it to knowledge. IBM has identified various steps which convert raw data into knowledge and wisdom, stating that the key is understanding relationships, patterns and principles

identifying the process as the DIKW model (IBM, 2018). Throughout this project, key notions have been discussed which have identified the significant data analytical progress made within elite sports and how it can be utilized within decision making. In particular, the conversion of raw data from the acclaimed SAP-HANA relational database to information through the SEMMA process yielded a predictive model which when correctly utilized enables organizations to extract useful knowledge and wisdom. Despite highlighting error, the constructed model enables organizations to identify which players show concrete numerical variables represented in their salary value. Through evaluating the predictive model an organization can determine the usefulness of the information and determine their decisions, thereby gaining context to the timely information, also known as knowledge. Lastly, with increasing time and data, organizations have the potential to develop further strategies in hopes of reaching optimal decisions which could entail the desired outcome of winning the Stanley cup. Through the process of data analytics, raw NHL data has been contextualized, ultimately having the potential for superior decision-making.

5.0 Discussion & Conclusion

This chapter will be split into two key segments (discussion and conclusion). The discussion segment will delve into the meaning, importance, and relevance of previous chapters. In particular, the discussion segment will explain and evaluate the importance of key findings due to performed data analysis. The conclusion segment of this chapter will conclude this project by focusing on overall interpretations and the importance behind this study. Furthermore, recommendations (what actions could be taken to further elaborate upon this project, and what further studies should follow) will be discussed.

5.1 Discussion

Throughout this project, a variety of research objectives have been discussed, and ultimately answered using literature, data modeling and data analysis. To address the usefulness of the overall project within the current NHL climate, the following will delve into the summary of key findings and how they have the potential to influence the decision-making process.

5.1.1 Future impact

Within the literature review chapter, it is clear that there is ongoing technological development within elite sports, with the continuous development of IOTSport and the use of relational databases in the form of SAP-HANA. Although rapid advances have been made, there still lies a gap within the field of data analytics and elite sports. Our research, focusing on developing a predictive model, has the potential to yield a direct impact on an organization's decision-making process; by analyzing player salaries, organizations can equate predicted values with real values and alter their decisions accordingly. Not only can the predictive model be utilized within the present state of elite sports, however, but it also holds the potential for further development, to maximize efficiency throughout all sporting decisions. The current situation has organizations spending upwards of $2 million per annum on scouts; Schuckers et al. research paper delves into the Toronto Maple Leafs, which have a total of 23 scouts listed on their respective webpages (Micheal E. Schuckers, 2015). Listed academic literature further illustrates that organizations should develop an ongoing data analysis department, to fully understand the power that data holds, as the potential for knowledge extraction to effective decision making. This project highlighted that through the SEMMA process a predictive model can accurately determine a player's salary, provided further resources an organization has the ability to develop data analytical models which could have a significant impact on long-term decision making.

5.1.2 Further Research

Predictive data mining is merely one realm of the expanding field of machine learning and data science. Though this project identifies how a predictive model, through supervised learning, can yield positive decision reassurance, there are several fields that ought to be delved into, to further deduce knowledge for even better decision making. Below are merely a few suggested research topics that have seen limited research.

5.1.2.1 Quantification of personality traits

The realm of elite sports psychology is continuously growing, as teams understand that a player in action performance is often determined by a player's personality and off-field actions. This field of study has seen limited academic research, yet there is a growing sense for sports psychologists, as denoted by the 'American psychological association' who have stated that there is a growing demand for sports psychologists as organizations aim to tackle issues of mental health (Wier, 2018). As noted within 'model limitations' within ice hockey it is critical to motivating players, hence why certain players are given a salary that relates not only to their on-ice performance, however their off-ice actions. With the evident need for psychological understanding within elite sports, a study could stem to quantify characteristic traits in hopes of understanding which players require further attention or which players could boost the overall motivation of a team's active roster.

5.1.2.2 Goaltender Analysis

Within this project, data utilized consisted merely of players within the positions LW, RW, C and D, however, there is a complex position in the form of the goaltender. Within the NHL and several other sports, the goaltender plays a significant role, and it is imperative that the correct player be chosen for the position. Within Ice hockey, each individual goaltender is statistically analyzed with a wide array of variables being identified, namely, save percentage, average goals per game or goals save above average. With the importance of the role, a further study examining goaltender data could provide a vast array of knowledge to organizations, as often within elite sports goaltenders are mistakenly assessed merely on a handful of statistical values which do not necessarily always convey the correct information.

5.2 Conclusion

This project provides a review of the literature on data mining within sports and performs data analysis through supervised learning to bridging the two domains. The continuous evolution of IOTs and the transcendence into the world of sports has seen the rise of new data sources stemming from a wide array

of sensors and other IOTSport. Furthermore, the utilization of SAP-HANA a relational database management system offers a main-memory-centric data management platform further facilitating an in-depth opportunity to attain and analyze big data.

A major example of how sports organizations could utilize data analytics is presented within this project, how the constructed predicted Neural Network model can be utilized to enhance the decision-making process. With a mere MAE of $870,000 the established model highlights significant utility, particularly within the realms of decision-making. Understanding players and their worth is a critical dimension within the NHL and other elite sports, by quantifying a player's performance an organization understands the purchasing power that a player can yield. Furthermore, through deep data analysis, not only do organizations understand which players are performing at an equal, higher, or lower level than their provided salary, but such information if leveraged effectively can lead to further knowledge ultimately mitigating poor decisions which the NHL has been seen too often. Furthermore, the continuous rise in data sources and improvements within technology, will merely increase the effectiveness of data mining techniques. The rise in data will mean organizations will have the ability to develop more complex models which can decrease model error (MAE) to points where they have all the necessary knowledge to make an optimal decision. This project highlights the power of third-party data, and its potential impact on player related decision making, with further tacit knowledge and private databases an organization could establish a competitive advantage through data analysis.

Even though this project discusses limitations, and room for further literature studies, this project provides a detailed analysis that can be used to further develop the notion of big data in elite sports, and how machine learning can impact decision making to improve upon the current system at play.

References

Agrawal, R. I. (1993). Database Mining: A Performance Perspective. *Transactions on Knowledge and Data Engineering*, 914-925.

Aisyah Mohd Noor, L. H. (2015). Big data: the challenge for small research groups in the era of cancer genomics. *British Journal of Cancer*, 1405-1412.

Akter S, W. S. (2016). How to improve firm performance using big data analytics capability and business strategy alignment? *International Journal of Prod Economics*, 113-131.

Amir Zadeh, D. T. (2020). Predicting Sports Injuries with Wearable Technology and data analysis. *Information Systems Frontiers*, 1-15.

Antonia Daderman, S. R. (2018). *Evaluating Frameworks for Implementing Machine Learning in Signal Processing: A comparative Study of CRISP-DM, SEMMA and KDD.* Stockholm: EXAMENSARBETE INOM TEKNIK.

Arora, J. S. (2004). Global Optimization Concepts and Methods for Optimum Design. *ScienceDirect*.

BDMS. (2014, August 20). *Big Data – An ultimate weapon in the field of sports?* Retrieved from Big Data made simple: https://bigdata-madesimple.com/big-data-an-ultimate-weapon-in-the-field-of-sports/

Bengio, J. B. (2012). Random Search for hyper-parameter optimization. *Journal of Machine Learning* , 281-305.

Bhageshpur, K. (2019). *Data Is The New Oil -- And That's A Good Thing.* Forbes. Retrieved from Data Is The New Oil -- And That's A Good Thing

Bharadwaj A, E. S. (2013). Digital business strategy: toward a next generation of insights. *MISQ*, 471-482.

Bojanova, I. (2014). IT Enhances Football at World Cup 2014. *IEEE*, 12-17.

Caparrós, T. C. (2018). Low external workloads are related to higher injury risk in professional male basketball games. *Journal of sports science & medicine*, 289-297.

CapFriendly. (2020). *CapFriendly*. Retrieved from CapFriendly: https://www.capfriendly.com/

CapFriendly. (2020). *CapFriendly*. Retrieved from Mario Lemieux: https://www.capfriendly.com/players/mario-lemieux

CK, D. (2014). Beyond data and analysis. *Commun ACM*, 39-41.

Clausen, H. (2012). *Important concepts of machine learning.* New Delhi: World Technologies.

Condor. (2020). *Linear Correlation*. Retrieved from Condor.edu: https://condor.depaul.edu/sjost/it223/documents/correlation.htm

Dietl, H. E. (2009). Governance of Professional Sports Leagues—Cooperatives versus Contracts. *International Review of Law and Economics*, 127-37.

EuropeanCommision. (2017). *The Internet of Things: reshaping the sport industry.* European Commision.

Falk, D. B. (1992). The Art of Contract Negotiation. *Marquette*, 1-29.

Fernando Santos, L. S. (2017). The Role of Team Captains in Integrating Positive Teammate Psychological Development in High-Performance Sport. *The sport Psychologist* , 1-11.

Franz Farber, N. M. (2015). The SAP HANA Database – An Architecture Overview. *IEEE*, 1-6.

Frost, J. (2020). *How to Interpret Adjusted R-Squared and Predicted R-Squared in Regression Analysis*. Retrieved from How to Interpret Adjusted R-Squared and Predicted R-Squared in Regression Analysis: https://statisticsbyjim.com/regression/interpret-adjusted-r-squared-predicted-r-squared-regression/

Gantz J, R. D. (2012). The digital universe in 2020: big data, bigger digital shadows, and biggest growth in the far east. *IDC*, 1-16.

Gartner. (2018). *Big Data*. Retrieved from Garnter.com: https://www.gartner.com/en/information-technology/glossary/big-data

Gujarati, D. N. (2003). *Basic Econometrics.* McGraw Hill.

Hourdin, F., Mauritsen, T., Gettelman, A., Golaz, J.-C., Balaji, V., Duan, Q., . . . Williamson, D. (2017). The Art and Science of Climate Model Tuning . *American Meteorological Society*, 589-602.

IBM. (2018, March 6). From data to knowledge. *IBM*.

J Anesthesiol, H. K. (2013). The prevention and handling of the missing data. *Korean Journal of Anesthesiology*, 402-406.

Jaime Sampaio, T. M.-G. (2015). Exploring Game Performance in the National Basketball Association Using Player Tracking Data. *Plos One*.

Jean-Marie Dufour, J. N. (2019). Grid Search. *ScienceDirect*.

Jeremiah Reyes, J. D. (2015). *Health Insurance Market.* ResearchGate.

Jiawei Han, J. P. (2012). *Max Normalization.* ScienceDirect.

Johan Muller, M. Y. (2019). Knowledge, power and powerful knowledge re-visited. *ResearchGate*, 1-19.

Jupyter. (2020). *Jupyter*. Retrieved from Jupyer: jupyter.org

Kalyani M Raval, K. (2012). Data Mining Techniques. *International Journal of Advanced Research in*, 4.

Kerr, Z. Y. (2015). College sports-related Injuries- United States. *Morbidity and Mortality Weekly Report*, 1330-1336.

Kwon O, L. N. (2014). Data quality management, data usage experience and acquisition intention of big data analytics. *International Journal of Information Management* , 387-394.

Lamba HS, D. S. (2015). Analysis of requirements for big data adoption to maximize IT business value in reliability. *2015 4th International conference on infocom technologies and optimization*, 1-6.

Lee, I. (2017). Big data: Dimensions, evolution, impacts, and challenges. *Business Horizons*, 293-303.

Levin, J. (2009). *Information on the 50-contract roster limit and 90-player maximum reserve list.* Nashville: National Hockey League.

Lopez, M. (2020, January 19). *THE EXTRA POINT*. Retrieved from Nfl: https://operations.nfl.com/stats-central/stats-articles/

MacDonald, D. (2018). *Hockey Success Is "A Game of Wise Decisions".* HockeyAdvisor.

Majumdar, B. G. (2019). Analysis and Detection of Diabetes Using Data Mining Techniques—A Big Data Application in Health Care. *Emerging Research in Computing, Information, Communication and Applications* , 443-455.

Memmert, R. R. (2016). Big data and tactical analysis in elite soccer: future challenges and opportunities for sports science. *SpringerPlus 5*.

Micheal E. Schuckers, S. A. (2015). You can beat the “market”: Estimating the return on investment for NHL team scouting. *Journal of Sports Analytics* , 111-119.

MIT. (2016). *MIT Technology Review*. Retrieved from technologyreview: https://www.technologyreview.com/s/600957/big-data-analysis

Müller O, J. I. (2016). Utilizing big data analytics for information systems research: challenges, promises and guidelines. *European Journal of Information systems*, 289-302.

N. Mansouri, M. M. (2019). Using data mining techniques to improve replica management in cloud computing. *Springer-Verlag*.

Nazanin Vafaei, R. A.-M. (2019). Normalization techniques for collaborative networks. *Emerald Insight*.

NBA. (2020). *NBA.com*. Retrieved from National Basketball Association: nba.com/stats

NHL. (2015, February 20). *NHL, SAP announce multi-year partnership; unveil new statistics platform that launches today*. Retrieved from NHL.com: https://www.nhl.com/news/nhl-sap-announce-multi-year-partnership-unveil-new-statistics-platform-that-launches-today/c-754248

NHL. (2020). *NHL*. Retrieved from National Hockey League: http://www.nhl.com/ice/page.htm?id=26366#:~:text=MINIMUM%20PLAYER%20SALARY&text=The%20minimum%20NHL%20player%20salary,is%20extended%20by%20the%20Union).

Norton, S. (2014). Germany’s 12th Man at the World Cup: Big Data. *WallStreet Journal*.

O'Brien, J. (2020, July 10). *NHL salary cap to stay flat at $81.5M; bad news for big spenders, free agents*. Retrieved from NBCSports: https://nhl.nbcsports.com/2020/07/10/nhl-salary-cap-to-stay-flat-at-81-5m-bad-news-for-big-spenders-free-agents/#:~:text=With%20the%20CBA%20extended%20through,the%202020%2D21%20NHL%20season.

P. Guillemin, a. P. (2009). Internet of things strategic research roadmap. *The Cluster of European Research Projects*.

P.P.Ray. (2014). Internet of things based physical activity monitoring (PAMIoT): an architectural framework to monitor physical activity. *CALCON*, 32-34.

Pääkkönen P, P. D. (2015). Reference architecture and classification of technologies, products and services for big data systems. . *Big Data Res 2*, 166-186.

Patrick Mikalef, I. O. (2017). Big data analytics capabilities: a systematic literature review and research agenda. *Information systems and e-Business management*, 547-578.

Patrick Ward, J. W. (2018). Business Intelligence: How Sport Scientists Can Support Organization Decision Making in Professional Sport. *International Journal of Sports Physiology and Performance* , 544-546.

Piatetsky-Shapiro, G., & Parker, G. (2018). *Lesson: Data Mining, and Knowledge Discovery: An introduction*. Retrieved from KDnuggets: https://www.kdnuggets.com/data_mining_course/x1-intro-to-data-mining-notes.html

Pincivero, D. M. (1997). *A physiological review of American football.* Springer International Publishing.

Qaiser, U. S. (2014). A Comparative Study of Data Mining Process Models (KDD, CRISP-DM and SEMMA). *International Journal of Innovation and Scientific Research*, 217-222.

Rafael G. Mantovani, A. L. (2015). Effectiveness of Random Search in. *IEEE*, 1-9.

Ray, P. P. (2014). *Home Health hub internet of things (H3IoT): an architectural framework for monitoring health of elderly people.* Chennai: ICSEMR.

Ray, P. P. (2015). *Internet of Things for Sports (IoTSport): An Architectural Framework for Sports and Recreational Activity.* Gangtok: Sikkim University .

Reitermanov´a, Z. (2010). *Data Splitting.* Prague: Charles University, Faculty of Mathematics and Physics.

Robert P. Schumaker, O. K. (2010). Sports Data Mining. In O. K. Robert P. Schumaker, *Sports Data Mining* (pp. 5-15). Springer US.

Rowley, J. (2007). The wisdom hierarchy: representations of the DIKW hierarchy. *Journal of Information Science*, 163-180.

Russom, P. (2011). Big Data analytics. *TDWI* , 1-35.

Ruth Dmonte, A. D. (2017). Big Data in Sports Leverage Big Data in Sports: An Insight using SAP HANA. *international journal of Engineering Research & Technology*.

Safaa Alkatheri, S. A. (2019). A Comparative Study of Big Data Frameworks. *International Journal of Computer Science and Information Security*.

Santos, A. A. (2008). KDD, SEMMA AND CRISP-DM: A PARALLEL OVERVIEW. *Scientific Repository of the polytechnic institute of Porto*, 1-10.

SAP. (2016, July 18). *VIDEO: The NBA Experience Starts With Live Data*. Retrieved from VIDEO: The NBA Experience Starts With Live Data: https://blogs.sap.com/2016/07/18/the-nba-experience-starts-with-live-data/

Schoenfeld, B. (2019). *How Data (and Some Breathtaking Soccer) Brought Liverpool to the Cusp of Glory.* New York: NY Times. Retrieved from https://www.nytimes.com/2019/05/22/magazine/soccer-data-liverpool.html

Schroeck M, S. R.-M. (2012). Analytics: The real-world use of big data. *IBM Global Business Services* , 1-20.

Seddon JJ, C. W. (2017). A model for unpacking big data analytics in high-frequency trading. *Journal of Business Resources* , 300-307.

Sherlock, H. (2019, September 1). *The average cost of a Premier League player and why we should all despair*. Retrieved from FootballFancast: https://www.footballfancast.com/premier-league/the-average-cost-of-a-premier-league-player-and-why-we-should-all-despair

Somayya Madakam, R. R. (2015). Internet of Things (IoT): A Literature Review. *Scientific Research publishing*, 164-173.

Statista. (2018). *Statista*. Retrieved from North America sports market size from 2009 to 2023 (in billion U.S. dollars)*: https://www.statista.com/statistics/214960/revenue-of-the-north-american-sports-market/

Statista. (2019). *National Hockey League franchise value by team in 2019*. Retrieved from Statista: https://www.statista.com/statistics/193732/franchise-value-of-national-hockey-league-teams-in-2010/

Statista. (2020, October). *Global digital population as of October 2020*. Retrieved from Global digital population as of October 2020: https://www.statista.com/statistics/617136/digital-population-worldwide/

StatSports. (2020). *StatSports*. Retrieved from StatSports: https://statsports.com/

Staudohar, P. (1998). Salary Caps in Professional Team Sports. *Compensation and Working conditions*, 3-11.

Steinberg, L. (2015). *CHANGING THE GAME: The Rise of Sports Analytics.* Forbes.

Stratos Idreos, F. G. (2012). MonetDB: Two Decades of Research in Column-oriented Database Architectures. *ResearchGate*.

Sumiran, K. (2018). An Overview of Data Mining Techniques and Their Application in Industrial Engineering . *Asian Journal of Applied Science and Technology*, 947-953.

TransferMarket. (2020). *Premier League market* . Retrieved from Premier League 20/21: https://www.transfermarkt.com/premier-league/startseite/wettbewerb/GB1

Udayan Khurana, H. S. (2017). *Feature Engineering for Predictive Modeling using Reinforcement Learning.* Cornell University.

Ugnar, D. (2012). *NHL: Why Captains Are More Important in Hockey Than Any Other Major Sport.* Bleacherreport.

Wagstaff, K. C. (2001). Constrained K-means Clustering with Background. *ln lcml*, 577-584.

Wang, L., & Jones, R. (2018). Big Data Analytics of Network Traffic and Attacks. *IEEE explore*.

White, C. (2011). *Using big data for smarter decision making IBM.* New York: Yorktown Heights.

Wier, K. (2018). *A growing demand for sport psychologists.* APA.

Woo, K. (2018). PAID TO PLAY: AN ANALYSIS OF NHL DEFENSEMEN SALARY IN RELATION TO INDIVIDUAL. *Journal of Sports Management*, 1-23.

Zachary Shelly, R. F. (2020). Using K-means Clustering to Create Training Groups for Elite American Football. *International Journal of Kinesiology & Sports Science*, 1-17.

Appendix

Predicted_Salaries

Player_ID	Salary	Salary_pred	Prediction difference	Average Error	Mean Absolute error
Corey Perry-2017	$8,625,000	$6,258,780	$2,366,220	$149,005	$873,030
Ryan Getzlaf-2017	$8,250,000	$8,490,574	($240,574)		
Derek Stepan-2017	$6,500,000	$5,266,382	$1,233,618		
Ryan Kesler-2017	$6,875,000	$4,009,322	$2,865,678		
Hampus Lindholm-2017	$5,250,000	$3,754,325	$1,495,675		
Oliver Ekman-Larsson-2017	$5,500,000	$3,690,545	$1,809,455		
Adam Henrique-2017	$4,000,000	$4,208,368	($208,368)		
Kevin Bieksa-2017	$4,000,000	$2,646,814	$1,353,186		
Ryan O'Reilly-2017	$7,500,000	$5,710,981	$1,789,019		
Alex Goligoski-2017	$5,475,000	$3,935,766	$1,539,234		
Niklas Hjalmarsson-2017	$4,100,000	$2,899,423	$1,200,577		
Cam Fowler-2017	$4,000,000	$5,156,818	($1,156,818)		
Rickard Rakell-2017	$3,789,444	$5,322,882	($1,533,438)		
Mark Giordano-2017	$6,750,000	$6,179,175	$570,825		
Kyle Okposo-2017	$6,000,000	$3,709,853	$2,290,147		
Jason Demers-2017	$3,937,500	$3,197,301	$740,199		
Jakob Silfverberg-2017	$3,750,000	$3,766,278	($16,278)		
Johnny Gaudreau-2017	$6,750,000	$5,652,103	$1,097,897		
Sean Monahan-2017	$6,375,000	$4,907,530	$1,467,470		
Jordan Staal-2017	$6,000,000	$5,911,239	$88,761		

Table 3: Predicted values snippet